Programa de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Guilherme Goldschmidt

ARTERIAL: Um Modelo Inteligente para a Prevenção ao Vazamento de Informações de Prontuários Eletrônicos utilizando Processamento de Linguagem Natural

São Leopoldo, 2021

Guilherme Goldschmidt

ARTERIAL: UM MODELO INTELIGENTE PARA A PREVENÇÃO AO VAZAMENTO DE INFORMAÇÕES DE PRONTUÁRIOS ELETRÔNICOS UTILIZANDO PROCESSAMENTO DE LINGUAGEM NATURAL

Dissertação apresentada como requisito parcial para a obtenção do título de Doutor pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Orientador:
Prof.Dr. Rodrigo da Rosa Righi

São Leopoldo
2021

G623a Goldschmidt, Guilherme.
Arterial : um modelo inteligente para a prevenção ao vazamento de informações de prontuários eletrônicos utilizando processamento de linguagem natural / Guilherme Goldschmidt. – 2021.
85 f. : il. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2021.
"Orientador: Prof. Dr. Rodrigo da Rosa Righi"

1. Prevenção ao vazamento de informação. 2. Processamento de linguagem natural. 3. Prontuários médicos eletrônicos. I. Título.

CDU 004

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Silvana Dornelles Studzinski – CRB 10/2524)

(Esta folha serve somente para guardar o lugar da verdadeira folha de aprovação, que é obtida após a defesa do trabalho. Este item é obrigatório, exceto no caso de TCCs.)

A minha família, meu orientador Rodrigo da Rosa Righi e ao amigo Felipe André Zeiser.

AGRADECIMENTOS

Este trabalho só foi possível devido à colaboração de várias pessoas em diferentes aspectos da pesquisa. Agradeço ao meu orientador Prof. Rodrigo da Rosa Righi, que me aceitou sob sua orientação como aluno de mestrado, compartilhando seus conhecimentos e tempo na condução dos trabalhos. Também por oferecer todo o suporte do qual precisei ao longo desse caminho tortuoso, sendo sempre atencioso e gentil. Agradeço também ao meu colega de curso e laboratório, Felipe André Zeiser, por me ajudar neste trabalho com feedbacks, incessantes incentivos e sempre preocupado com meu bem estar. Muito obrigado.

Um agradecimento especial à minha família pelo apoio emocional e financeiro ao longo da minha vida. Ao meu pai, Ilvo, e à minha mãe, Aneli, por serem minha inspiração para a vida. À minha esposa, Sarah, por todo carinho, cuidados, suporte emocional e apoio extra ao longo dos meus estudos. Ao meu irmão, Pedro Henrique, pela amizade e pelos momentos de diversão.

Agradeço também aos excelentes professores do PPGCA (Programa de Pós-Graduação em Computação Aplicada) da Unisinos. Da mesma forma, gostaria de agradecer às secretárias e funcionários da Unisinos pela ajuda e paciência durante o mestrado. Também gostaria de expressar meus agradecimentos aos meus colegas de laboratório, Fabiano, Felipe, João, Luis e Régis, pelos momentos de diversão , café e sugestões de trabalho e vida. Sem vocês, aquele laboratório não é o mesmo.

Por fim, esse trabalho não seria possível sem o apoio financeiro de várias agências financiadoras. Gostaria de agradecer ao CNPq pela bolsa de mestrado.

RESUMO

Na última década, houve um aumento constante de violações de segurança na área de saúde. Um estudo sobre privacidade de pacientes e segurança de dados mostrou que 94% dos hospitais tiveram pelo menos uma violação de segurança nos últimos dois anos. Na maioria dos casos, os ataques tiveram origem por parte de atores internos. Dessa forma, é essencial que as organizações de saúde protejam suas informações sensíveis como, resultados de exames, diagnósticos, prescrições, pesquisas e informações pessoais de clientes. Um vazamento de dados sensíveis pode resultar em uma grande perda econômica e ou dano à imagem da organização. Há no Brasil ainda a Lei Geral de Proteção de Dados Pessoais (LGPD), que dispõem sobre diversos aspectos da proteção pessoal de informações. Sistemas para a proteção da informação foram se concretizando ao longo dos últimos anos, como *firewalls*, *intrusion detection and prevention systems* (IDS/IPS) e *virtual private networks*. No entanto, essas tecnologias funcionam muito bem em dados bem definidos, estruturados e constantes, diferente do que são os prontuários médicos que possuem campos de escrita livre. Para complementar essas tecnologias há os sistemas de prevenção ao vazamento de dados, denominados *Data Leakage Prevention Systems* (DLPS). Sistemas de DLP ajudam a identificar, monitorar, proteger e reduzir os riscos de vazamento de dados sensíveis. No entanto as soluções convencionais de DLP utilizam apenas comparações por assinatura e ou comparação estática. Dessa forma, propomos desenvolver um modelo com base em novas tecnologias como Processamento de Linguagem Natural (PLN), Reconhecimento de Entidades (NER) e Redes Neurais Artificiais (RNA) com o objetivo de ser mais assertivo na extração de informação e no reconhecimento de entidades. Contribuindo assim com novas perspectivas à literatura e por conseguinte à comunidade cientifica. Foram implementadas e testadas três abordagens, duas a partir de RNA e a seguinte com base em algoritmos de aprendizado de máquina. Como resultado, a abordagem que teve em sua implementação a utilização de algoritmo de aprendizado de máquina atingiu 98.0% de Precisão, 86.0% de Recall e 91.0% de F1-Score.

Palavras-chave: Prontuários Médicos Eletrônicos. Prevenção ao Vazamento de Informação. Processamento de Linguagem Natural.

ABSTRACT

Over the past decade, there has been a steady increase in healthcare security breaches. A study on patient privacy and data security showed that 94% of hospitals had at least one security breach in the past two years. In most cases, the attacks originated from internal actors. Therefore, it is essential that healthcare organizations protect their sensitive information such as test results, diagnoses, prescriptions, surveys, and personal customer information. A leak of sensitive data can result in a great economic loss and/or damage to the organization's image. There is also in Brazil the General Law for the Protection of Personal Data (LGPD), which provides for various aspects of the personal protection of information. Information protection systems have been taking shape over the last few years, such as firewalls, intrusion detection and prevention systems (IDS/IPS) and virtual private networks (VPN). However, these technologies work very well on well-defined, structured and constant data, unlike medical records that have free writing fields. Complementing these technologies are Data Leakage Prevention Systems (DLPS). DLP systems help to identify, monitor, protect and reduce the risk of leaking sensitive data. However, conventional DLP solutions use only subscription comparisons and/or static comparisons. Thus, we propose to develop a model based on new technologies such as Natural Language Processing (NLP), Entity Recognition (NER) and Artificial Neural Networks (ANN) to be more assertive in extracting information and recognizing entities. Thus contributing with new perspectives to literature and therefore to the scientific community. Three approaches were implemented and tested, two based on ANN and the next based on machine learning algorithms. As a result, the approach that took in its implementation the use of machine learning algorithm reached 98.0% of Accuracy, 86.0% of Recall and 91.0% of F1-Score.

Keywords: Electronic Health Record. Data Leakage Prevention. Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

CE	Critério de Exclusão
CI	Critério de Inclusão
DLP	*Data Leakage Prevention*
DLPS	*Data Leakage Prevention Sistem*
RES	Registro Eletrônico de Saúde
FP	Falso Positivo
HIPAA	Lei de Portabilidade e Responsabilidade do Seguro de Saúde
HITECH	Lei de Tecnologia da Informação em Saúde para Saúde Econômica e Clínica
IDS	*Intrusion Detection System*
IPS	*Intrusion Prevention System*
K-NN	*K-nearest Neighbors*
LSTM	*Long Short-term Memory*
ML	*Machine Learning*
NCB	*National Center for Biomedical Computing*
NER	*Named-entity Recognition*
NIH	*National Institutes of Health*
NLTK	*Natural Language Toolkit*
PLN	Processamento de Linguagem Natural
POS	*Part of Speech*
QP	Questão de Pesquisa
RNA	Redes Neurais Artificiais
RNR	Redes Neurais Recorrentes
RSL	Revisão Sistemática da Literatura
TF	*Term Frequency*
TP	*True Positive*
TsF	*Time Series Forecast*
T-Tree	T-Árvore Binária
URL	*Uniform Resource Locator*
VPN	*Virtual Private Network*

SUMÁRIO

1 INTRODUÇÃO

A disseminação da informação e os meios pelos quais ela se encontra disponível têm aumentado em grande escala. As organizações estruturam suas informações de forma que elas estejam acessíveis em todos os meios e plataformas disponíveis. Ainda, incentivam o trabalho colaborativo visando a eficiência e a produtividade.

Essa disseminação da informação, contudo, gera um ambiente vulnerável, uma vez que os atores internos são considerados responsáveis por 62.4% dos vazamentos de informações (INFOWATCH, 2019). A Figura 1 exemplifica os tipos de violadores organizados por ano. Quando trouxemos esse cenário para o ambiente hospitalar/clínico o problema se intensifica. Organizações hospitalares devem seguir leis extremamente rígidas quanto a privacidade dos dados. Como exemplo, há a Lei de Portabilidade e Responsabilidade do Seguro de Saúde (HIPAA) e a Lei de Tecnologia da Informação em Saúde para Saúde Econômica e Clínica (HITECH). Ambas as leis são baseadas na legislação estadunidense e aplicam multas por violação ou com base na quantidade de registros afetados (ACT, 1996; SERVICES et al., 2011).

No Brasil, há a Lei Geral de Proteção de Dados Pessoais (LGPD) com aprovação no ano de 2018 e com vigor a partir de agosto de 2020 (BRASIL, 2018). A LGPD dispõem sobre diversos aspectos da proteção pessoal de informações, incluindo assim pacientes. Em relação a saúde, a lei destaca três princípios principais quanto aos dados: a finalidade, adequação e necessidade. A finalidade discute quanto a motivação do uso do dado, a adequação trata da finalidade e tratamentos compatíveis e a necessidade aborda quanto a utilização somente dos dados necessários para a execução da finalidade (GONÇALVES, 2021). Assim, os usuários terão o direito de saber a finalidade, por meio de quem e quando suas informações serão utilizadas, além de delimitar a possibilidade de acesso aos dados.

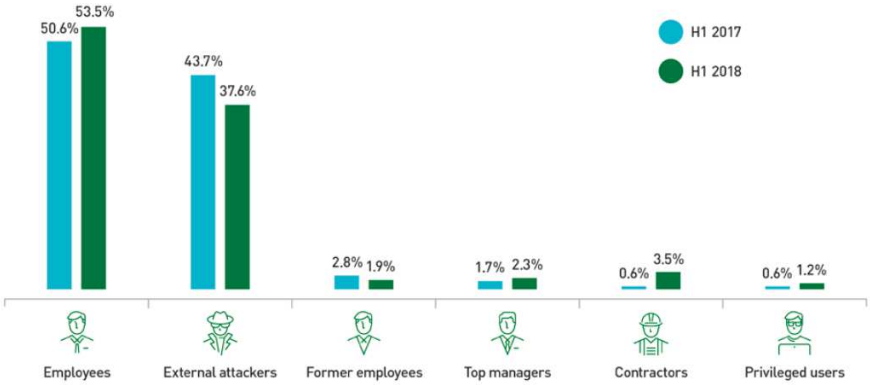
Em paralelo com a aplicação de novas regulamentações aos dados de pacientes, temos a crescente utilização do Registros Eletrônicos de Saúde (RES) que tem disponibilizado uma quantidade impressionante de informações em formato digital (ZHANG; ELHADAD, 2013). O RES pode conter vários grupos de dados, como alergias, sinais vitais, evoluções médicas, consultas médicas, resultados de exames laboratoriais, imagens médicas e diagnósticos (ISO/TR:14639-2, 2014).

1.1 Motivação

Na última década, houve um aumento constante de violações de segurança na área de saúde (PATIL; SESHADRI, 2014). Em 2013, a *Kaiser Permanente* notificou seus pacientes que suas informações de saúde foram comprometidas devido ao roubo de uma unidade flash USB não criptografada contendo registros de pacientes (MCCANN, 2013).

O relatório de investigação de violação de dados da *Verizon* declarou que sua divisão de

Figura 1 – Tipos de violadores por ano, 2017 e 2018



Fonte: Infowatch (2019)

investigação e segurança forense compilou dados de incidentes de segurança relatados e encontrou 621 violações de dados confirmadas (TEAM et al., 2013). Além disso, um estudo sobre privacidade de pacientes e segurança de dados mostrou que 94% dos hospitais tiveram pelo menos uma violação de segurança nos últimos dois anos (PONEMON, 2011). Na maioria dos casos, os ataques tiveram origem por parte de atores internos (PATIL; SESHADRI, 2014). A Figura 2 informa os números de vazamentos de dados de 2006 a 2018.

Ademais as violações, com o vigor da LGPD no Brasil os hospitais/clínicas devem tratar os dados dos pacientes de acordo com a nova regulamentação. Assim, fazendo com que os dados sejam garantidos quanto a utilização adequada, ou seja, somente dos dados necessários para o cumprimento da finalidade. Também há previsto na lei penalizações diversas como, por exemplo, multa de até 2% do faturamento institucional, limitadas a R$50 milhões (BRASIL, 2018).

Como visto anteriormente, é essencial que as organizações de saúde protejam suas informações sensíveis, resultados de exames, diagnósticos, prescrições, pesquisas e informações pessoais de clientes. Um vazamento de dados privados pode resultar em uma grande perda econômica e ou dano à imagem da organização.

Sistemas para a proteção de informação foram se concretizando ao longo dos últimos anos, como *firewalls*, sistemas de detecção e prevenção de intrusão (IDS/IPS) e redes virtuais privadas (VPN). No entanto, estas tecnologias funcionam muito bem em dados bem definidos, estruturados e constantes. Para complementar essas tecnologias existe o sistema de prevenção ao vazamento de dados, do inglês *Data Leakage Prevention (DLPS)*. Sistemas de DLP ajudam a identificar, monitorar, proteger e reduzir os riscos de vazamento de dados sensíveis. Soluções DLP são utilizadas para detectar e impedir que usuários não autorizados obtenham dados

Figura 2 – Quantidade de vazamentos agrupados por ano

Fonte: Infowatch (2019)

confidenciais e até para proteger dados que podem ser compartilhados acidentalmente (SULLIVAN, 2008; SHABTAI et al., 2012).

Como apresentado anteriormente, DLPSs, diferente de outras tecnologias voltadas para a proteção de dados, são projetados com a capacidade de analisar dados não estruturados. Esses sistemas de prevenção ao vazamento de dados, tem como características realizar análises de conteúdo e de contexto (TAHBOUB; SALEH, 2014). A análise de conteúdo é definida como a melhor abordagem para dados não estruturados (SECUROSIS, 2010). No entanto, dos estudos relacionados a presente proposta, apenas 03 utilizam essa abordagem (THORLEUCHTER; POEL, 2012; DESHPANDE et al., 2015; MASHECHKIN et al., 2015). Destes, nenhum aborda o contexto clínico/hospitalar. Além disso, há os que fazem uso de dados sintéticos para seu desenvolvimento (DESHPANDE et al., 2015).

Dos trabalhos relacionados ao tema proposto neste estudo, não foram encontradas soluções que apresentassem a aplicação de redes neurais artificiais, nem a utilização de semântica para a detecção da transformação de textos. Uma revisão sistemática aplicada a DLPs apresenta como fragilidade destes sistemas a não detecção da transformação dos dados (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). Ainda, embora os trabalhos relacionados apresentem a aplicação de técnicas de mineração de texto e a utilização de algoritmos de aprendizado de máquina, pesquisas que realizaram a comparação entre algoritmos de aprendizado de máquina e redes neurais artificiais aplicadas a textos, mostraram uma melhor performance das redes neurais artificiais (VINODHINI; CHANDRASEKARAN, 2016). Dessa forma, realizar experimentos a fim de comparar as abordagens se mostra válido.

1.2 Questão de Pesquisa

A questão de pesquisa que o modelo proposto busca responder é a seguinte: *Como seria um modelo de prevenção ao vazamento de informações voltado ao contexto hospitalar, que faça uma análise de conteúdo utilizando processamento de linguagem natural?*

Como definição, DLPSs são sistemas analíticos designados para proteger dados de divulgação não autorizada em todos os estados, usando ações corretivas desencadeadas por um conjunto de regras (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). Esta definição contém três atributos principais que distinguem os DLPSs das medidas de segurança convencionais. Primeiro, os DLPSs têm a capacidade de analisar o conteúdo de dados confidenciais e o contexto. Segundo, os DLPSs podem ser implantados para fornecer proteção de dados confidenciais em diferentes estados da informação, ou seja, em trânsito, em uso e em repouso. O terceiro atributo é a capacidade de proteger os dados por meio de várias ações corretivas, como notificação, auditoria, bloqueio, criptografia e quarentena. A proteção normalmente começa com a capacidade de detectar possíveis vazamentos por meio de heurísticas, regras, padrões e impressões digitais (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016).

Sistemas DLP costumam comparar conteúdos com base em regras e expressões regulares (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). A correspondência de dados não estruturados é realizada pelos sistemas existentes computando e armazenando hashes unidirecionais para conteúdo protegidos, e rastreando possíveis vazamentos identificando conteúdo semelhante em outros documentos. No entanto, a deficiência de usar o hash unidirecional para a correspondência não estruturada de dados é que essa abordagem funciona apenas se uma cópia exata dos dados for transferida; não é eficaz na detecção de situações em que uma versão alterada, reformulada (por exemplo, usando palavras sinônimas ou de código) ou resumida dos dados originais vazou.

Além disso, a detecção com base na correspondência de expressões regulares pode ser evitada facilmente, pois um usuário mal-intencionado pode remover habilmente dos dados todas as palavras-chave problemáticas. São considerados pontos fracos das abordagens atuais de análise de conteúdo dos DLPS a utilização de expressões regulares ou baseadas em regras, correspondência exata de arquivos, correspondência parcial de documentos e análise estatística (SECUROSIS, 2010).

1.3 Objetivos

Autores em estudos sobre DLPS, sugerem em suas conclusões a utilização de técnicas de classificação de texto como trabalhos futuros (RAMAN; KAYACIK; SOMAYAJI, 2011; ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). Então, como forma de aprimorar sistemas de DLP, a utilização de processamento semântico das informações é uma

alternativa inteligente. Em paralelo, esforços significativos foram dedicados à criação de terminologias e bases de conhecimento padrão. Facilitando então a extração de informações e o raciocínio sobre dados brutos. Agora então, o gargalo do processamento de informações médicas, portanto, mudou de onde coletar dados e recursos para como usar os recursos de conhecimento e criar modelos escaláveis para processar grandes quantidades de texto. Como muitos dos dados são registrados de forma narrativa e não estruturada, como em notas clínicas e publicações biomédicas, a qualidade das ferramentas básicas de Processamento de Linguagem Natural (PLN) tem um impacto crítico no desempenho de tarefas de nível superior, como recuperação de informações, extração de informações e descoberta de conhecimento (ZHANG; ELHADAD, 2013).

Para lidar com esses problemas, um sistema que leve em consideração o Processamento de Linguagem Natural (PLN), analisando o conteúdo e utilizando Reconhecimento de Entidades Nomeadas (NER) ou Redes Neurais Artificiais (RNA) pode obter resultados mais precisos na detecção de alterações em conteúdo. Uma vez que esse sistema leva em consideração a semântica da linguagem e o aprendizado de máquina. Exemplos de sistemas que utilizam NER incluem extração de informações clínicas de relatórios de radiologia (FRIEDMAN et al., 1994; HRIPCSAK et al., 1995; FISZMAN et al., 2000) identificação de doenças e nomes de medicamentos em resumos de alta (CHAPMAN et al., 2001; MELTON; HRIPCSAK, 2005; UZUNER et al., 2011).

Esse trabalho possui como objetivo geral: *Desenvolver um modelo de **DLP** voltado ao **contexto clínico/hospitalar** que seja capaz de identificar conteúdos sensíveis com o objetivo de prevenir o vazamento de informação*, fazendo uso de **PLN** e **Machine Learning**. Será realizada também uma comparação entre diferentes abordagens na utilização de NER, a fim de adicionar ao modelo a abordagem com melhor desempenho.

Para atingir os objetivos gerais, foram definidos os seguintes objetivos específicos:

(i) Realizar um levantamento bibliográfica dos conceitos e tecnologias necessárias ao estudo;

(ii) Pesquisar e analisar estudos relacionados ao tema de prevenção a perda de dados;

(iii) Definir a base de dados a ser utilizada como referência para o modelo;

(iv) Baseando-se na análise realizada, desenvolver um modelo de DLP que agregue aos modelos já propostos encontrados na análise dos estudos relacionados;

(v) Desenvolver e treinar o modelo para a realização de testes;

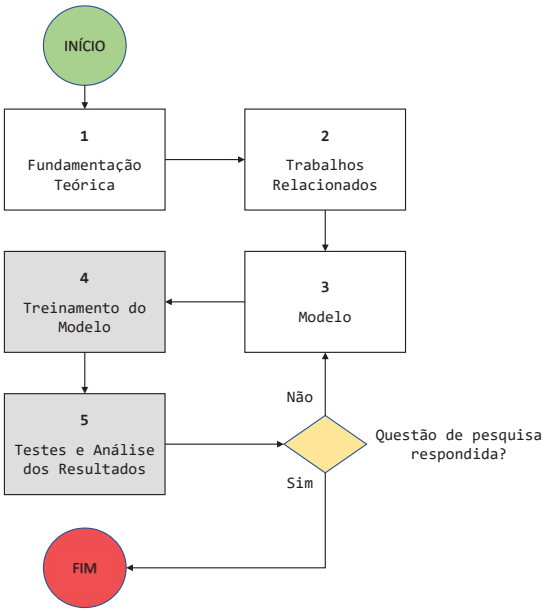
(vi) Realizar testes para a avaliação do modelo proposto;

(vii) Comparar os resultados obtidos com os trabalhos relacionados;

1.4 Etapas de Desenvolvimento da Pesquisa

O desenvolvimento da pesquisa está dividido em 5 etapas principais, são elas: (1) Fundamentação teórica; (2) Trabalhos relacionados; (3) Modelo; (4) Treinamento do modelo; (5) Testes e análise dos resultados. Inicialmente é realizado o estudo dos conceitos envolvidos no tema da pesquisa para formar o referencial teórico. Após isso, inicia-se a etapa de pesquisa de trabalhos relacionados ao tema da pesquisa. Ao final dessa etapa, é realizada uma análise com o objetivo de identificar as lacunas presentes no estado da arte. Assim, na terceira etapa, é proposto e desenvolvido o modelo com foco em atender os objetivos apresentados e responder à questão de pesquisa. Em seguida, são realizadas os treinamentos necessários ao modelo e na quinta etapa são realizados os testes e as análises dos resultados obtidos. A Figura 3 demonstra o fluxo de realização da pesquisa.

Figura 3 – Fluxograma das etapas de desenvolvimento da pesquisa. As caixas em cinza não estão finalizadas.



Fonte: Elaborado pelo autor.

1.5 Organização do Texto

O estudo é organizado de forma que o Capítulo 2 resume os principais conceitos e tecnologias que apoiam a proposta. O Capítulo 3 explica os trabalhos relacionados mais significativos, seus objetivos, métodos e resultados. Já o Capítulo 4 apresenta e detalha o modelo proposto. Os procedimentos realizados durante os experimentos do modelo proposto estão descritos no Capítulo 5. Os dados de avaliação e discussão dos resultados estão no Capítulo 6. Por fim, o Capítulo 7 apresenta as conclusões dos autores, destaca as contribuições e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A seção atual descreve conceitos fundamentais para o entendimento do modelo proposto. São apresentadas as definições e características de Sistemas de Prevenção a Perda de Dados e outros conceitos relevantes. Outras tecnologias como PLN e RNA são abordadas de forma que seus principais conceitos estão dispostos no texto.

2.1 Sistemas de Prevenção ao Vazamento de Dados

DLP é um acrônimo para *Data Loss Prevention* (OUELLET; MCMILLAN, 2013) e é reconhecido como uma área na qual são estudadas formas de prevenção a perda de dados (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). De acordo com Alneyadi, Sithirasenan e Muthukkumarasamy (2016) o termo DLP é definido como um vazamento de dados (ou perda de dados) e sua aplicabilidade no campo da Segurança da Informação é descrever divulgações indesejadas de informações. Em diferentes áreas o acrônimo DLP é referencia a diferentes termos como Prevenção a Perda de Dados, Prevenção ao Vazamento de Dados e Prevenção a Perda/Vazamento de Informações (SECUROSIS, 2010). Neste estudo o acrônimo DLP fará referência a *Data Leakage Prevention* (RAMAN; KAYACIK; SOMAYAJI, 2011; TAHBOUB; SALEH, 2014) traduzida nesse estudo para Prevenção ao Vazamento de Informações.

Como forma de implementar DLP são utilizados os DLPS. DLPSs são sistemas que permitem a aplicação dinâmica de política com base na classificação do conteúdo durante uma operação. Um DLPS descreve um conjunto de tecnologias e técnicas de inspeção usadas para classificar o conteúdo das informações contidas em um objeto - como um arquivo, e-mail, pacote, aplicativo ou armazenamento de dados - enquanto estiver em repouso (no armazenamento), em uso (durante uma operação) ou em trânsito (em uma rede); e a capacidade de aplicar dinamicamente uma política - como registrar, relatar, classificar, realocar, marcar e criptografar - e/ou aplicar proteções de gerenciamento de direitos de dados corporativos (OUELLET; MCMILLAN, 2013).

2.1.1 Estados da Informação

Sistemas de DLP classificam as informações em três estados que incluem dados ’em trânsito’, ’em uso’ e ’em repouso’(OUELLET; MCMILLAN, 2013; ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016; SULLIVAN, 2008). Dados em trânsito são os dados transmitidos de um nó para outro. Esse tipo de dado viaja internamente entre nós na mesma rede ou externamente entre nós que pertencem a redes diferentes. Dados em uso são os dados acessíveis ao usuário na forma de documentos, e-mails e aplicativos, dados que estejam em interação com o usuário (TAHBOUB; SALEH, 2014). Esse tipo de

dado aparece em texto simples, para que possa ser facilmente interpretado e processado. Dados em repouso é o tipo de dados armazenados nos repositórios. Consiste em bancos de dados de aplicativos, arquivos de backup e sistemas de arquivos. Normalmente, é protegido por fortes controles de acesso, incluindo mecanismos físicos e lógicos (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). Os estados da informação apresentados são descritos na Figura 4.

Figura 4 – Estados da Informação

Fonte: Adaptado de Alneyadi, Sithirasenan e Muthukkumarasamy (2016)

2.1.2 Análise de Contexto

A informação é um conjunto de dados que representa um ponto de vista. Um dado não tem valor antes de ser processado, a partir do seu processamento, ele passa a ser considerado uma informação, que pode gerar conhecimento (STANDARDIZATION, 2005). Para realizar essa contextualização de um dado em informação, se faz necessária uma interpretação pertinente. Podemos elencar duas formas de se analisar dados a fim de compreende-los: Análise de Contexto e Análise de Conteúdo.

Para a análise de contexto os dados são tratados de uma forma menos profunda. Busca-se uma análise macro, levando em consideração uma quantidade grande de dados e suas relações. Por exemplo, se um usuário estiver enviando dados para outra entidade, serão estudados atributos contextuais como origem, destino, tempo, tamanho e formato. Esses atributos podem ser usados para formar padrões de processo ou transação e, com base em políticas predefinidas, é possível identificar discrepâncias (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). Assim, é possível ter uma perspectiva mais abrangente da informação, no entanto especificidades de dados podem revelar informações relevantes ao cenário analisado.

A análise contextual geralmente fornece o contexto para as políticas de análise de conteúdo (SECUROSIS, 2010). Soluções mais assertivas costumam fazer o uso de ambas as análises, de contexto e de conteúdo. Visto que, além de dados de atributos como endereços, formatos e tamanhos o conteúdo é também fator relevante para a geração da informação. O contexto é altamente útil e deve ser incluído como parte de uma solução geral (SECUROSIS, 2010).

2.1.3 Análise de Conteúdo

Uma forma de exemplificar essa abordagem é pensar no conteúdo como uma carta e no contexto como o envelope. O contexto incluiria origem, destino, tamanho, destinatários, remetente, informações do cabeçalho e qualquer outra coisa que não seja o conteúdo da própria carta. Assim, mesmo analisando os dados e gerando informação, o entendimento da comunicação ainda é deficitário. O primeiro passo na análise de conteúdo é receber o envelope e abri-lo. A análise de conteúdo será capaz de verificar o conteúdo escrito na carta. Dessa forma, sendo possível relacioná-lo também ao contexto. Assim, podemos dizer que a Análise de Conteúdo busca ser mais profunda quanto a complexidade dos dados (SECUROSIS, 2010).

Como a análise de conteúdo envolve a verificação dentro dos campos de dados e a análise do próprio conteúdo, sua vantagem é que, mesmo usando o contexto, não somos restringidos por ele (SECUROSIS, 2010). Caso se faça necessário proteger um formulário sensível, o ideal é protegê-lo por completo - não apenas campos notoriamente sensíveis. O objetivo é proteger os dados, não o envelope, por isso abrir a carta, ler e decidir como lidar com isso é mais sensato. Isso é mais difícil e demorado do que a análise de contexto (SECUROSIS, 2010).

Como o principal objetivo do uso de DLPSs é a proteção de dados confidenciais, é mais importante focar no conteúdo em si do que no contexto (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). A análise de conteúdo nos DLPSs é feita através de três técnicas principais: impressão digital de dados (incluindo correspondência exata ou parcial), expressão regular (incluindo correspondência baseada em dicionário) e análise estatística (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016).

2.2 Mineração de Texto

A mineração de texto é um processo para extrair conhecimento útil de fontes de dados, embora semelhante à mineração de dados, existem divergências. As fontes de dados usadas no processo de mineração de texto são formadas por coleções de documentos e o conhecimento é extraído de dados não estruturados presentes nos documentos (FELDMAN; SANGER et al., 2007). Os dados não estruturados analisados pela mineração de texto geralmente são textos escritos em uma linguagem natural (JO, 2018). O processo também pode ser aplicado para extrair conhecimento de dados semi-estruturados, como arquivos HTML e XML (YANG; CHEN, 2002).

A aplicação do processo de mineração de texto para adquirir conhecimento pode se diferenciar de acordo com o tipo de conhecimento pretendido. A mineração de texto abrange tarefas que implementam uma variação do processo da mineração de dados, permitindo a extração de diferentes informações de fontes de dados não estruturadas. Algumas tarefas compreendidas pela mineração de texto são recuperação de informações, extração de informações, sumarização, classificação e análise sentimental (GUPTA; LEHAL et al., 2009; WEIGUO et al., 2005; ALLAHYARI et al., 2017).

Independentemente da tarefa de mineração de texto, o estágio inicial do processo é a execução de uma ou mais técnicas de pré-processamento. Existem várias técnicas de pré-processamento e cada uma delas opera uma melhoria distinta na preparação de dados não estruturados. O pré-processamento é responsável por transformar os dados de texto bruto em um formato intermediário, o que pode melhorar a extração de padrões (MUNKOVÁ; MUNK; VOZÁR, 2013; ANGIANI et al., 2016; UYSAL; GUNAL, 2014).

Geralmente uma técnica única de pré-processamento não é suficiente para preparar o texto para uma tarefa de mineração de texto, dessa forma é bastante comum que diferentes técnicas sejam implementadas em sequência (FELDMAN; SANGER et al., 2007). Através de técnicas de pré-processamento como a *tokenização*, é possível realizar uma segmentação de um texto. Então, é possível realizar uma limpeza através da exclusão de caracteres como emojis, pontuação, e a remoção de palavras de parada (*stop words*). Podemos ainda padronizar as palavras aplicando um processo de derivação ou lematização ou ainda analisar gramaticalmente usando uma marcação de parte do discurso (*Part-Of-Speech*). É possível também, atribuir pesos a cada palavra em um documento para destacar seu grau de importância usando esquemas de ponderação de termos (JO, 2018; WEISS et al., 2010; KANNAN; GURUSAMY, 2014).

2.2.1 Processamento de Linguagem Natural

A linguagem natural é inerente a textos não estruturados, onde há a necessidade de algum pré-processamento de documentos a fim de realizar uma forma de estruturação dos dados (RAJMAN; VESELY, 2004). Já o PLN é um subcampo da ciência da computação, inteligência artificial e linguística, que visa a compreensão da linguagem natural usando computadores (LIDDY, 2001; MANNING; MANNING; SCHÜTZE, 1999; ZHAI; MASSUNG, 2016). Muitos dos algoritmos de mineração de texto fazem uso extensivo de técnicas de PLN, como parte da marcação do discurso (PoS), análise sintática e outros tipos de análise linguística (ALLAHYARI et al., 2017).

No entanto, os analisadores modernos consistem basicamente de sete etapas principais: tokenização, segmentação de sentença, marcação de parte do discurso (PoS), lematização, NER e análise sintática (SLANKAS et al., 2014; LIDDY, 2001). **Tokenização**, considerada a primeira etapa do processamento do texto, tem como objetivo basicamente a conversão de

documentos de textos em pedaços (ZHAI; MASSUNG, 2016; ALLAHYARI et al., 2017). Para atingir o objetivo, a tokenização detecta palavras, pontuações e outros itens de textos (SLANKAS et al., 2014; ALLAHYARI et al., 2017). **Segmentação de sentença** identifica a construção de frases e possibilita então a segmentação das sentenças em tokens (SLANKAS et al., 2014).

PoS identifica e atribui marcações baseadas em gramática, como por exemplo substantivo, verbos e adjetivos (SLANKAS et al., 2014). PoS é realizada por ferramentas de PLN e pode ser considerada uma tarefa complicada, com ferramentas auxiliares disponíveis para auxiliar no processo (RINGGER et al., 2007).

A **Lematização** considera a análise morfológica das palavras, ou seja, agrupando as várias formas flexionadas de uma palavra para que possam ser analisadas como um único item (ALLAHYARI et al., 2017). Em outras palavras esse processo gera a palavra raiz comum para um grupo de palavras relacionadas. Por exemplo, cantei, cantaria e cantar são todas as formas de um lema comum de “canta” (SLANKAS et al., 2014).

Reconhecimento de entidade nomeada (NER) procura identificar palavras ou frases relacionadas com uma entidade em um conteúdo não estruturado e classificá-las de acordo com grupos previamente definidos, como pessoas, organizações, locais ou horários (AGGARWAL; ZHAI, 2012a; JURAFSKY, 2000). As arquiteturas NER usam uma combinação de regras de alta precisão, correspondência probabilística e técnicas de aprendizado de máquina (JURAFSKY, 2000). Como resultado uma entidade nomeada consegue identificar alguma entidade do mundo real, por exemplo "Google Inc", "Estados Unidos"e "Barack Obama"e atribuir seus significados como, por exemplo, "Organização", "País"e "Pessoa"respectivamente (ALLAHYARI et al., 2017).

Ao realizar o reconhecimento de entidades nomeadas deve-se ter cuidado, pois dicionários podem não conter todas as formas de entidades nomeadas de um determinado tipo de entidades. Além disso, as entidades nomeadas são frequentemente dependentes do contexto, por exemplo, *Big Apple* pode ser a fruta ou uma forma de referência a cidade de Nova York (ALLAHYARI et al., 2017). Portanto, para essa tarefa existem dicionários específicos que acrescem as taxa de acerto no reconhecimento das entidades no contexto clínico (CHEN et al., 2015). As técnicas utilizadas para o reconhecimento de entidades nomeadas são classificadas em três grupos (GOYAL; GUPTA; KUMAR, 2018):

- Baseado em regras: sistemas normalmente baseados na elaboração manual de regras que fazem uso de listas previamente cadastradas para a identificação e classificação de entidades. Apesar de serem considerados muito eficientes, tais sistemas são dependentes de pessoas com conhecimento sobre a área analisada para a definição das regras e listas de entidades;

- Baseado em aprendizado: estes sistemas utilizam aprendizado de máquina, que permite o aprendizado automatizado de padrões e sequências que podem ser aplicados nas tomadas

de decisão. Estes sistemas não necessitam da intervenção de pessoas para a elaboração de regras. Este tipo de sistema pode fazer uso de aprendizado supervisionado, semi-supervisionado ou não supervisionado;

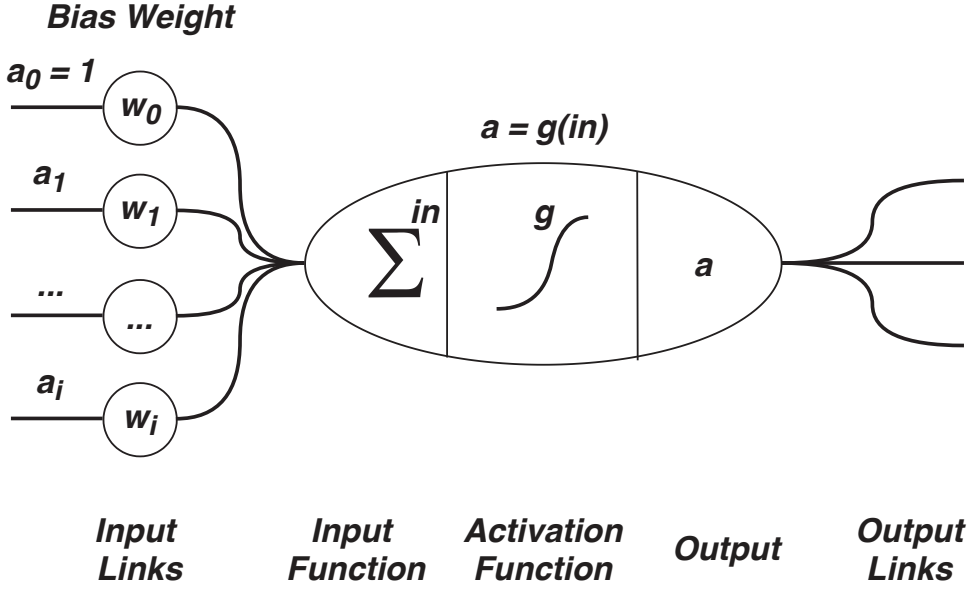
- Sistema híbrido: implementa técnicas dos dois sistemas anteriores, produzindo um resultado baseado no uso múltiplas técnicas de aprendizado de máquina ou de técnicas de aprendizado de máquinas juntamente com regras criadas manualmente.

2.3 Redes Neurais Artificiais

Na última década, a utilização de RNA para solução de diversos problemas se tornou foco das linhas de pesquisa. Contudo, o primeiro estudo a propor um neurônio artificial foi desenvolvido ainda na década de 1940 (MCCULLOCH; PITTS, 1943). Este neurônio artificial é inspirado no funcionamento do cérebro humano, buscando simular o funcionamento das sinapses realizadas pelo sistema nervoso humano e fornecer ao computador a capacidade de aprendizado artificial (COPPIN, 2010).

O fluxo de processamento das informações em um neurônio artificial, representado na Figura 5, é similar ao de um neurônio biológico (BUDUMA; LOCASCIO, 2017). Os sinais de entrada (a_i) são multiplicados pelos pesos $(w_{i,j})$ associados a cada sinal de entrada, para então passarem por uma função de entrada, que tipicamente corresponde ao somatório de todos os sinais ponderados. A saída dessa passa por uma função de ativação, que corresponde ao efeito que a entrada exerce na definição do próximo neurônio (COPPIN, 2010).

Figura 5 – Neurônio Artificial



Fonte: Russell e Norvig (2013)

Portanto, matematicamente um neurônio artificial pode ser representado pela seguinte equação:

$$f(x) = \sigma\left(\sum_{i=1}^{n} x_i w_i + b\right) \tag{2.1}$$

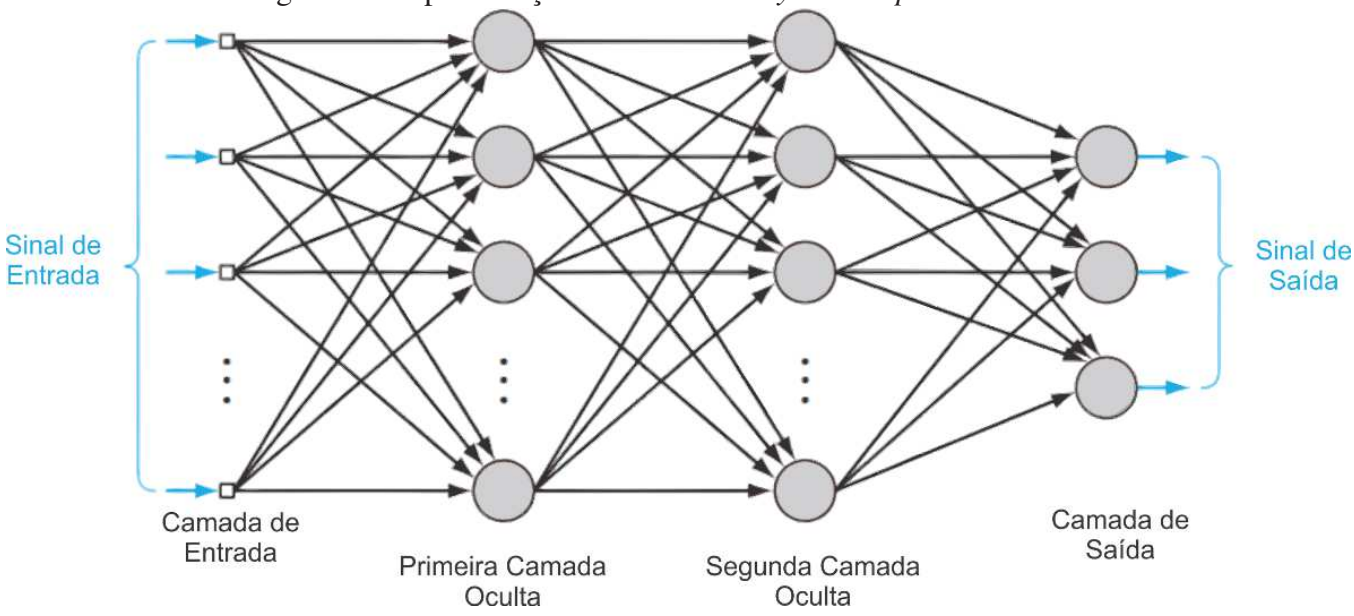
Onde x_i é a entrada i, w_i é o peso, correspondente a sinapse do neurônio, adicionado a entrada i, b corresponde ao bias e σ que é a função de ativação (RUSSELL; NORVIG, 2013).

2.3.1 Multilayer perceptron

A utilização de um neurônio é capaz de resolver problemas simples, por exemplo, a representação de portas lógicas E e OU (RUSSELL; NORVIG, 2013; COPPIN, 2010), porém para problemas mais complexos e que exigem maior capacidade de generalização necessita-se utilizar um conjunto de neurônios, comumente organizados em camadas. Uma arquitetura de RNAs multicamada capaz de modelar funções complexas é a *Multilayer Perceptron*.

Uma *Multilayer Perceptron* pode ser dividida em três componentes: a camada de entrada, uma ou mais camadas ocultas e a camada de saída. Cada sinal de saída de um neurônio é repassado a todos os neurônios da camada adjacente, esse processo é conhecido como *propagation*. O sinal de saída de uma camada é dado pelo processamento de todos os sinais de saída da camada anterior (COPPIN, 2010), conforme pode ser observado na Figura 6.

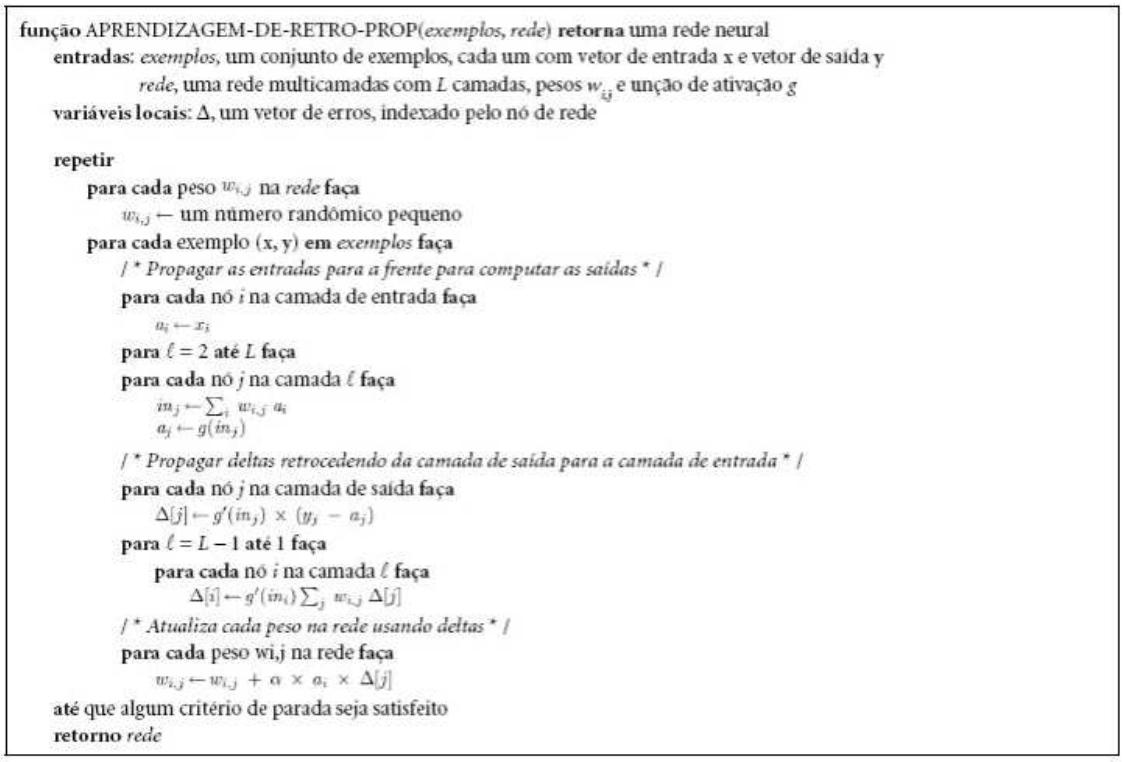
Figura 6 – Representação de uma *Multilayer Perceptron.*



Fonte: Haykin (2009)

O aprendizado em uma rede *Multilayer Perceptron* ocorre mensurando o erro da previsão com o conjunto de dados rotulados e propagando o erro para todos os neurônios da rede utilizando um algoritmo conhecido como *backpropagation* (HAYKIN, 2009). Na Figura 7 é apresentado um pseudoalgoritmo da técnica de *backpropagation*.

Figura 7 – Pseudoalgoritmo da técnica de *backpropagation*.



função APRENDIZAGEM-DE-RETRO-PROP(*exemplos*, *rede*) **retorna** uma rede neural
entradas: *exemplos*, um conjunto de exemplos, cada um com vetor de entrada x e vetor de saída y
rede, uma rede multicamadas com L camadas, pesos $w_{i,j}$ e unção de ativação g
variáveis locais: Δ, um vetor de erros, indexado pelo nó de rede

repetir
para cada peso $w_{i,j}$ na *rede* **faça**
$w_{i,j} \leftarrow$ um número randômico pequeno
para cada exemplo (x, y) **em** *exemplos* **faça**
/ * *Propagar as entradas para a frente para computar as saídas* * /
para cada nó i na camada de entrada **faça**
$a_i \leftarrow x_i$
para $\ell = 2$ **até** L **faça**
para cada nó j na camada ℓ **faça**
$in_j \leftarrow \sum_i w_{i,j}\, a_i$
$a_j \leftarrow g(in_j)$
/ * *Propagar deltas retrocedendo da camada de saída para a camada de entrada* * /
para cada nó j na camada de saída **faça**
$\Delta[j] \leftarrow g'(in_j) \times (y_j - a_j)$
para $\ell = L - 1$ **até** 1 **faça**
para cada nó i na camada ℓ **faça**
$\Delta[i] \leftarrow g'(in_i) \sum_j w_{i,j}\, \Delta[j]$
/ * *Atualiza cada peso na rede usando deltas* * /
para cada peso wi,j na rede **faça**
$w_{i,j} \leftarrow w_{i,j} + \alpha \times a_i \times \Delta[j]$
até que algum critério de parada seja satisfeito
retorno *rede*

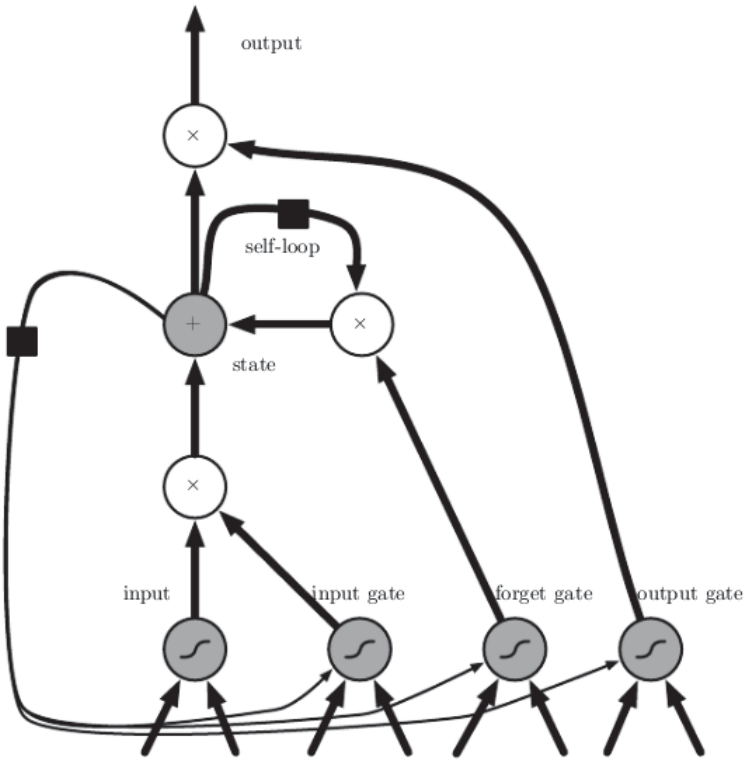
Fonte: Russell e Norvig (2013)

2.3.2 Long short-term memory

As *Long Short-Term Memory* (LSTM) são uma variante de Rede Neural Recorrente. As Redes Neurais Recorrentes se diferem das RNAs pela capacidade de armazenar informações de entradas processadas anteriormente. Esta característica tornam as redes baseadas em unidades LSTM relevantes quando aplicadas a problemas que possuem interdependência temporal, como para o processamento de vídeo e texto (CHOLLET, 2018).

Na LSTM os neurônios, ou células são conectadas recorrentemente uma a outra. Uma unidade da LSTM (Figura 8) é composta por quatro neurônios: *input*, *input gate*, *forget gate*, e *output gate*. O neurônio de *input*, computa a informação da mesma forma que um neurônio artificial. Já no *input gate*, a informação atual leva em consideração os estados anteriores armazenados no bloco de memória *state*. O *forget gate*, exerce a função da remoção de informações do bloco de memória *state*. Por fim, o *output gate* é responsável pela decisão de qual informação a unidade LSTM irá transmitir para as próximas câmadas (GOODFELLOW; BENGIO; COURVILLE, 2017).

Figura 8 – Diagrama de uma unidade da Rede Neural Recorrente LSTM.



Fonte: Goodfellow, Bengio e Courville (2017)

2.4 Contexto Hospitalar

As informações de um paciente são essenciais para seu tratamento, bem estar e para questões contábeis. Como forma de organizar essas informações as instituições de saúde fazem o uso de prontuários. Os prontuários são uma coleção de informações vitalícias sobre o paciente. Com o tempo, o prontuário médico do paciente acumula informações pessoais significativas, incluindo identificação, histórico de diagnóstico médico, renderizações digitais de imagens médicas, tratamentos, histórico de medicamentos, hábitos alimentares, preferência sexual, informações genéticas, perfis psicológicos, histórico de emprego, renda e subjetividade médica avaliações de personalidade e estado mental (MERCURI, 2004).

Com a evolução da tecnologia, o prontuário se tornou eletrônico, com diversos padrões e regulamentos (LOZANO-RUBÍ et al., 2016; MILSTEIN; BLANKART, 2016). O RES contém informações relevantes para o bem-estar, a saúde e o cuidado de um indivíduo, em forma processável por computador e representada de acordo com um modelo de informação padronizado (ISO/TR:14639-2, 2014). RES refere-se a uma estrutura de maneira eletrônica dos registros de saúde do paciente, coletada e armazenada em um repositório, que pode ser compartilhada por diferentes formatos digitais. O RES pode conter vários grupos de dados, como alergias, sinais vitais, consultas médicas, resultados de exames laboratoriais, imagens

médicas e diagnósticos (ISO/TR:14639-2, 2014; HEART; BEN-ASSULI; SHABTAI, 2017).

Os registros médicos de um paciente também são compartilhados com instituições de pagamentos, como empresas seguradoras, para justificar o pagamento dos serviços prestados. Os provedores de assistência médica também usam registros para gerenciar suas operações e melhorar a qualidade do serviço (APPARI; JOHNSON, 2010). Assim, mecanismos de proteção de segurança atuais precisam ser aprimorados para a proteção de registros, mas para manter a privacidade os níveis de segurança não devem ficar tão rígidos que os registros de saúde sejam inutilizáveis (ARCHER et al., 2011). Um modelo de DLP pode verificar a informação, classifica-la e acionar medidas de restrição, caso necessário, sem a necessidade da imposição de medidas que deterioram a usabilidade dos sistemas por parte do usuário.

2.5 Considerações Parciais

Esse capítulo abordou os principais conceitos envolvidos no tema proposto nesse estudo. Foram apresentados conceitos relacionados a mineração de textos, RNA e DLPS. Esses conceitos e suas características foram abordados a fim de oferecer um embasamento sobre as tecnologias. A mineração de texto e as RNAs são meios utilizados para a aplicação de PLN e para a classificação de conteúdo. Esses conceitos são pertinentes para um melhor entendimento de como será estruturado o modelo proposto. Ainda, foi apresentado o conceito de prontuários eletrônicos, esse conceito é pertinente para o entendimento do contexto ao qual o estudo está direcionado. O modelo apresentado nesse estudo está inserido dentro do contexto das tecnologias, modelos e técnicas descritas nesse capítulo.

3 TRABALHOS RELACIONADOS

Este capítulo tem como objetivo apresentar e discutir os estudos relacionados ao tema do modelo proposto. Este capítulo está organizado em (a) metodologia de pesquisa e (b) análise dos trabalhos. A busca inicial sobre o tema foi realizada a partir da aplicação de uma metodologia de pesquisa que consiste de um método de busca estruturado em bases de conhecimento como Springer, Scopus, entre outros.

3.1 Metodologia de Pesquisa

Esse estudo aplicou a metodologia da Revisão Sistemática de Literatura (RSL) (PAI et al., 2004). Essa metodologia consiste na aplicação de um método para extrair, sintetizar e resumir informações com base em uma coleção de pesquisas. Uma RSL é um meio para identificar, avaliar e interpretar toda a pesquisa disponível relevante para uma questão de pesquisa, área temática ou fenômeno de interesse específico (KITCHENHAM, 2004). Esse estudo foi conduzido de acordo com a estrutura fornecida por (KITCHENHAM; BUDGEN; BRERETON, 2015), que está organizada em três fases: (a) a Revisão do Plano, relacionada ao design do protocolo de pesquisa, compreendendo tarefas como as questões de pesquisa, estratégias de busca, e os critérios para selecionar os estudos; (b) Realizar a Revisão, envolvendo as tarefas de identificação do corpus a ser utilizado no estudo e que tipo de informação precisa ser extraída; (c) Revisão de documentos, compreendendo o relatório escrito, formatado como um trabalho de pesquisa, e que deve mostrar a contribuição para a área pesquisada.

3.1.1 Planejamento da Revisão

O primeiro passo relacionado ao protocolo RSL é determinar as questões de pesquisa, pois elas orientam a seleção de estudos e definem que tipo de dados precisam ser extraídos. Essa etapa é baseada no objetivo estabelecido para a RSL que pode ser refinado em várias perguntas. O objetivo desse trabalho é identificar diferentes aplicações de mineração de texto para resolver problemas relacionados à segurança cibernética. Com base nisso, esse trabalho determina as seguintes questões de pesquisa:

- QP1: Como seria a taxonomia representando as diversas aplicações da mineração de texto na área de segurança cibernética?

- QP2: Em quais contextos, a cibersegurança tem aplicado a mineração de texto?

- QP3: Quais estratégias de mineração de texto estão sendo utilizadas?

– QP4: Como vem sendo o desempenho da classificação de texto no domínio da segurança cibernética?

– QP5: Como o setor de segurança cibernética tem aplicado a mineração de texto em soluções do mundo real?

A QP1 tem como objetivo enumerar várias áreas de segurança cibernética que usam algumas tarefas de mineração de texto para lidar com problemas de segurança. Para organizar a taxonomia, os autores definiram diferentes domínios relacionados à segurança cibernética, e a taxonomia define quais atividades relacionadas aos domínios de segurança cibernética aplicam a mineração de texto. Além da questão de pesquisa anterior, a QP2 propõe apresentar os diferentes tipos de conteúdo analisados pelas atividades de cibersegurança relacionadas a cada domínio na taxonomia. O contexto mencionado na QP2 também considera os setores e tecnologias direcionados a aplicação de mineração de texto. Por fim, são avaliadas as linguagens analisadas pelos estudos presentes no corpus.

A QP3 apresenta as estratégias de uso da mineração de texto. A estratégia compreende três aspectos relacionados à aplicação de tarefas de *text mining*. O primeiro aspecto analisa quais tarefas de mineração de texto são aplicadas isoladamente. A QP3 também investiga quais tarefas foram combinadas para avaliar o desempenho da mineração de texto no domínio da segurança cibernética. Por fim, a questão avalia o uso de RNA para apoiar a aplicação de mineração de texto no domínio da segurança cibernética. A QP4 avalia o desempenho associado ao uso da classificação de textos nas diferentes atividades identificadas através da leitura do corpus da revisão. A última questão de pesquisa (QP5) investiga o uso de mineração de texto pela indústria de segurança cibernética, dando exemplos de aplicações que implementam mineração de texto no mundo real.

O protocolo também envolve a estratégia de busca utilizada para descobrir os estudos relacionados às questões de pesquisa. Essa RSL define uma estratégia ampla, com o objetivo de abranger, tanto quanto possível, as pesquisas relevantes associadas a ela. As palavras-chave são divididas em dois tópicos de pesquisa relacionados às principais áreas do estudo. As palavras-chave relacionadas à área de segurança cibernética compreendem as possíveis variações do termo. Também foram selecionados estudos relacionados a segurança da informação, porque há muito em comum entre os conceitos. Outras palavras-chave tentam cobrir as diferentes expressões vistas nos estudos para descrever o uso de algumas tarefas de mineração de texto. A Tabela 1 apresenta os dois grupos de palavras-chave. A *string* de busca foi desenvolvida utilizando o operador lógico "OR"entre todas as palavras-chave do mesmo grupo e o operador lógico "AND"entre os dois grupos.

A estratégia de pesquisa limita o uso de *string* de busca em alguns bancos de dados científicos de conhecida relevância no campo da computação. As fontes foram escolhidas com base em um estudo de referência (BRERETON et al., 2007) e outras RSLs conduzidas em ciência da computação (TRAN; ZDUN et al., 2017; MOUSTAKA; VAKALI;

Tabela 1 – Palavras-chave usadas para buscar estudos relevantes

Tópicos de Busca	**Palavras-chave**
Cybersecurity	cybersecurity, cybersecurity, cyber-security, information security
Text Mining	text mining, information retrieval, natural language processing, information extraction, text analysis, text classification, text clustering, text summarization

ANTHOPOULOS, 2018). Os bancos de dados utilizados nesse estudo são ACM[1], IEEE Xplore[2], Science Direct[3], Scopus[4], Springer[5] e Web of Science[6]. O protocolo especifica que a *string* de busca deve ser comparada com os campos título, palavras-chave e resumo em cada banco de dados. A pesquisa levou em consideração os estudos publicados nos últimos dez anos.

O planejamento também estabelece os critérios para excluir ou incluir um estudo nessa RSL. De acordo com (KITCHENHAM; BUDGEN; BRERETON, 2015), os critérios são essenciais para obter evidências de que os estudos contribuem para responder às questões da pesquisa. Esse estudo define seis critérios de exclusão (CE) e quatro critérios de inclusão (CI) mostrados na Tabela 2. Depois de executar as consultas de pesquisa nos diferentes bancos de dados, a primeira tarefa é reunir todos os resultados em um único arquivo e excluir entradas duplicadas. Esses estudos são considerados os trabalhos candidatos, e o próximo passo é a aplicação dos critérios de exclusão.

O primeiro critério de exclusão visa remover toda a literatura cinza encontrada no resultado, uma vez que foi decidido utilizar apenas estudos revisados por pares. O segundo critério remove os artigos curtos e o terceiro exclui os artigos secundários. O quarto critério mantém apenas trabalhos escritos em inglês e o quinto contém apenas a versão completa de cada trabalho candidato. O último critério de exclusão exclui todos os documentos candidatos cujos autores não podem acessar diretamente dos bancos de dados escolhidos na estratégia de pesquisa.

O protocolo do estudo definiu quatro critérios de inclusão para analisar trabalhos candidatos relacionados à aplicação de mineração de texto na área de segurança cibernética e para garantir que a mineração de texto seja usada para extrair conhecimento de conteúdo não estruturado. O primeiro critério limita os trabalhos candidatos relacionados a estudos que aplicam alguma técnica de mineração de texto para resolver um problema de segurança cibernética. Esse critério permite considerar exclusivamente trabalhos focados em segurança

[1] https://dlnext.acm.org/
[2] https://ieeexplore.ieee.org/
[3] https://www.sciencedirect.com/
[4] https://www.scopus.com/
[5] https://link.springer.com/
[6] https://webofknowledge.com

Tabela 2 – Critérios aplicados para incluir ou excluir estudos na RSL

ID	Critérios de Inclusão/Exclusão
CI1	Estudo está relacionado à aplicação de mineração de texto para solucionar um desafio de segurança cibernética
CI2	A mineração de texto deve ser o principal método usado para solucionar um desafio de segurança cibernética
CI3	As técnicas de mineração de texto devem ser aplicadas em conteúdo não estruturado como um documento
CI4	O estudo deve ter um experimento ou estudo de caso e apresentar resultados consistentes
CE1	Estudo não publicado em revista ou conferência
CE2	O estudo tem menos de oito páginas
CE3	O estudo é uma revisão de literatura
CE4	O estudo não é em inglês
CE5	Estudo é uma versão mais curta de outro estudo
CE6	Os autores não podem acessar o artigo completo

cibernética e remover estudos gerais que apenas realizam uma avaliação usando dados de segurança. No entanto, o objetivo não está associado à área. O próximo critério restringe a seleção a documentos candidatos que aplicam a mineração de texto como o método principal para obter os resultados, e esse critério é necessário porque um experimento pode aplicar vários métodos e a mineração de texto pode ter uma função secundária no resultado final.

O terceiro critério limita aos trabalhos candidatos que aplicam a mineração de texto para extrair conhecimento de conteúdo não estruturado, uma vez que alguns estudos usam métodos de mineração de texto para extrair conhecimento de conteúdo estruturado. Por exemplo, estudos que aplicam mineração de texto em logs, URLs e senhas não são incluídos na RSL. O último critério de inclusão mantém estudos que apresentam uma apresentação abrangente e coerente dos resultados com base em algum experimento ou estudo de caso.

O protocolo determinou a aplicação dos critérios de inclusão em três rodadas, além de alguns critérios de exclusão. A primeira rodada compreendeu a leitura do título e resumo. Na primeira rodada, foram removidos apenas os artigos que evidentemente não estavam relacionados às questões de pesquisa. A segunda rodada compreendeu uma análise através da leitura da introdução e das seções associadas à metodologia e ao experimento/estudo de caso. A segunda rodada proporcionou confiança para selecionar apenas documentos com as informações que precisavam ser extraídas para a terceira rodada. A última rodada compreendeu a leitura completa e a extração das informações para responder às questões da pesquisa.

Outra decisão tomada durante a fase de planejamento é sobre a qualidade dos estudos. A avaliação da qualidade é uma parte essencial do processo, porque esta etapa pode melhorar o valor da RSL, oferecendo aos revisores a opção de excluir trabalhos que não atinjam o nível de qualidade esperado (KITCHENHAM; BUDGEN; BRERETON, 2015). A RSL utilizou a

pontuação h5-index[7] como critério de qualidade, optando pela exclusão de todos os trabalhos publicados em procedimentos ou periódicos com pontuação menor que 10.

3.1.2 Conduzindo a Revisão

A segunda fase determina a implementação do protocolo, seguindo as decisões tomadas anteriormente. As consultas de pesquisa foram realizadas nos bancos de dados entre 12 de abril de 2019 e 17 de abril de 2019 e encontraram 2.472 trabalhos candidatos. Conforme definido no protocolo, a pesquisa buscou a expressão no título, palavras-chave e resumo. A única exceção foi a *Springer* que não possui essa opção. Nesse caso, a pesquisa buscou a expressão em todo o artigo. A etapa seguinte destacou 276 estudos com duas ou mais ocorrências nos resultados e removeu todas as duplicatas. Alguns estudos duplicados foram encontrados em diferentes bancos de dados, citando anos distintos de publicação. Nessas situações, foi decidido remover a entrada com o ano de publicação mais antigo.

Após a exclusão de estudos duplicados, os restantes 2.196 trabalhos candidatos foram aplicados aos critérios de exclusão. Na primeira rodada, os critérios de exclusão removeram 929 trabalhos candidatos, e foram analisados o resumo e o título dos 1.267 remanescentes. Após a análise desses campos, 309 estudos foram selecionados para uma análise mais aprofundada no segundo turno. Os 784 estudos candidatos não foram selecionados porque, a partir da leitura do título e do resumo, ficou evidente que eles não estavam relacionados à segurança cibernética ou à mineração de texto. A Figura 9 apresenta as etapas executadas para selecionar os estudos. Na figura, duas caixas são exibidas juntamente com alguns critérios, onde a caixa cinza, à esquerda, informa o número de estudos selecionados para avançar na seleção, e a caixa azul, à direita, apresenta o número de estudos removidos com base em o critério.

Para concluir a segunda rodada, foi necessário fazer o download dos demais documentos candidatos, verificar se eles não correspondiam a alguns critérios de exclusão e ler as seções definidas no protocolo. O primeiro e o segundo critérios não foram aplicados porque não excluiriam estudos adicionais. Considerando o conjunto inicial de 309 estudos, analisamos apenas 269, uma vez que 40 estudos foram removidos por corresponderem aos critérios de exclusão. Após a análise de cada trabalho candidato, foram selecionados 113 estudos para continuar na próxima etapa do processo.

Foi então decidido aplicar a avaliação da qualidade na segunda rodada. Essa decisão foi tomada devido ao número de trabalhos candidatos. O protocolo estabeleceu o uso do índice h5 como um limiar de qualidade, e não há razão para uma leitura detalhada de um conjunto de artigos que o critério de qualidade removerá. A aplicação de todos os periódicos e conferências para verificar a métrica do índice h5 foi realizada de 28 a 30 de junho de 2019 e resultou na exclusão de 26 estudos; portanto, o número de estudos que chegavam à próxima rodada eram

[7] https://scholar.google.com/intl/en/scholar/metrics.html#metrics

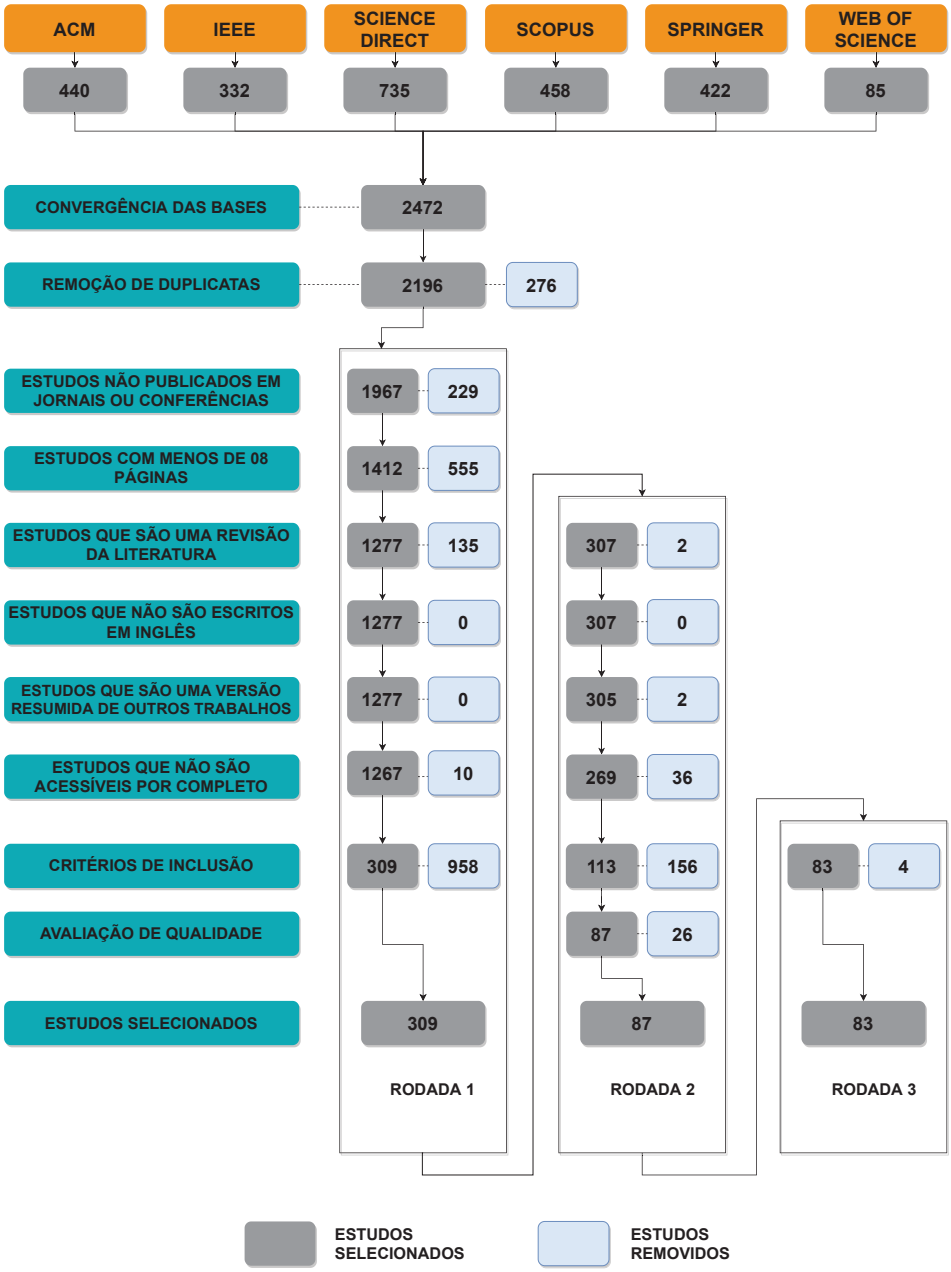


Figura 9 – Processo para selecionar os estudos incluídos na RSL

87. Após a leitura completa, decidiu-se por excluir quatro estudos, restando assim 83 artigos.

3.2 Análise do Corpus

Esta seção tem como objetivo a discussão dos trabalhos selecionados como resultado da aplicação da revisão sistemática. A seção contém uma análise dos trabalhos selecionados, seus objetivos, técnicas utilizadas, modelos de aplicação, contextos e resultados.

Ao final do processo da revisão sistemática foram selecionados um total de 83 estudos que

Tabela 3 – Estudos que utilizam Processamento de Texto

Autor	Título	Objetivo	Classificação	Sumarização	Clusterização	Content Analysis	PLN	Redes Neurais Artificiais	Contexto
(ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013)	Adaptable N-gram classification model for data leakage prevention	Classificação da Informação	Sim	Não	Não	Não	Não	Não	Documentos Eletrônicos
(HUANG et al., 2018)	A novel mechanism for fast detection of transformed data leakage	Prevenção	Sim	Não	Sim	Não	Não	Não	Geral
(GONZALEZ-COMPEAN et al., 2019)	A policy-based containerized filter for secure information sharing in organizational environments	Automatização de Controles	Não	Não	Sim	Não	Não	Não	Cloud Computing
(MASHECHKIN et al., 2015)	Applying text mining methods for data loss prevention	Predição de Vazamento	Não	Não	Não	Sim	Não	Não	Correio Eletrônico
(KATZ; ELOVICI; SHAPIRA, 2014)	CoBAn: A context based model for data leakage prevention	Prevenção	Não	Não	Sim	Não	Não	Não	Documentos Eletrônicos
(WANG; JIN, 2011)	Data leakage mitigation for discretionary access control in collaboration clouds	Prevenção	Não	Não	Não	Não	Não	Não	Cloud Computing
(ZARDARI; JUNG, 2016)	Data security rules/regulations based classification of file data using TsF-kNN algorithm	Classificação da Informação	Sim	Não	Não	Não	Não	Não	Geral
(THORLEUCHTER; POEL, 2012)	Improved multilevel security with latent semantic indexing	Classificação da Informação	Sim	Não	Não	Sim	Não	Não	Geral
(DESHPANDE et al., 2015)	The Mask of ZoRRo: preventing information leakage from documents	Prevenção	Não	Não	Não	Sim	Não	Não	Documentos Eletrônicos

Fonte: Elaborada pela autor

foram classificados e organizados em uma taxonomia. Para o estudo em questão, serão levados em consideração 09 artigos, pois estes foram classificados em um nicho dentro da segurança cibernética nomeado DLP. Estes estudos, que buscam responder a questões de segurança cibernética e mineração de texto apontadas ao início desse capítulo, ajudarão no desenvolvimento do modelo Arterial. A Tabela 3 apresenta os estudos selecionados. Os estudos estão organizados em ordem alfabética por título e relacionados de acordo com as técnicas de mineração de texto utilizadas, quanto a aplicação de redes neurais artificiais e contexto de aplicação.

3.2.1 Processamento de Linguagem Natural

Alguns DLPSs usam padrões e assinaturas para comparação entre dados, a fim de detectar o vazamento de informação. A detecção ocorre quando esses padrões e assinaturas são correspondidos ou quando um alto grau de similaridade é observado (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016; WANG; JIN, 2011; HUANG et al., 2018). No entanto, os dados confidenciais nem sempre são enviados em sua forma original. De fato, dados confidenciais podem ser expostos a muitos tipos de modificações. Por exemplo, os usuários podem editar perceptivamente documentos confidenciais adicionando, subtraindo ou substituindo linhas ou até parágrafos antes de enviá-los. Além disso, a semântica de um documento pode ser reescrita na forma de resumos ou elaborações longas. Essas variações podem alterar a identidade do documento original; portanto, padrões de dados e assinaturas se tornam ineficazes (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016).

Soluções para a detecção de informação pertinente e detecção de alteração de conteúdo são

propostas. Huang et al. propõem um mecanismo para detecção rápida de vazamento de dados transformados. No estudo é utilizando um método adaptativo de aprendizado com grafos ponderados e um *score walk* para realizar essa verificação (HUANG et al., 2018). No entanto, as abordagens propostas utilizam métricas de avaliação que não levam em consideração as variações linguísticas. Fazem uso de contagem e palavras, análise de frequência e adição de pesos (DESHPANDE et al., 2015; GONZALEZ-COMPEAN et al., 2019).

Shu and Yao propõem dividir um documento em partes menores e calcular o valor do *hash* de cada parte como uma melhor abordagem ao problema apresentado pela comparação de *hashes* analisando o contexto (SHU; YAO, 2012). A variedade linguística de um idioma possibilita a transformação do texto, sem alteração de conteúdo. Dessa forma, essa abordagem é excludente e pouco eficiente visto que uma simples alteração no texto pode torna-lo inacessível. Um Sistema de DLP efetivo deve ter a capacidade de classificar dados confidenciais semanticamente (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016).

É possível verificar que embora estudos proponham a utilização de técnicas de mineração de texto com o objetivo de detectar a transformação de conteúdo, nenhum dos estudos faz uso do processamento de linguagem natural. Recursos de PLN como a lematização que considera a análise morfológica das palavras é essencial para detectar a transformação do texto (ALLAHYARI et al., 2017). Dessa forma, é possível concluir que devido a grande quantidade de recursos linguísticos que um idioma oferece, uma análise que não leva em consideração a semântica de um texto, não faz uso de recursos importantes que podem oferecer uma melhor acurácia na detecção de informações.

3.2.2 Redes Neurais Artificiais

Com o objetivo de atribuir categorias de segurança e restrições a documentos, ou seja, classificá-los, os autores Thorleuchter and Poel propõem uma estratégia de pré-processamento dos documentos. Esse pré-processamento é realizado com base no uso de *Latent semantic analysis (LSA)* para reduzir a matriz gerada pelos documentos e *Logistic Regression* para classificação (THORLEUCHTER; POEL, 2012).

Zarrdari and Jung descrevem seu modelo como um classificador baseado em regras e regulamentos de segurança de dados de arquivos através da utilização de *TsF-kNN* (ZARDARI; JUNG, 2016). *TsF-kNN* é uma implementação do algoritmo de classificação *k-Nearest Neighbors (K-NN)* em combinação com *Term Frequency (TF)* que realiza a contagem de termos em um determinado arquivo. As etapas definidas pelos autores são: treinamento da base de dados usando o modelo *T-Tree*, pré-processamento (*tokenização* e *stop words*) e aplicação do classificador (ZARDARI; JUNG, 2016).

Os estudos apresentados na Tabela 3 descrevem o uso de técnicas de mineração de texto como classificação (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013;

HUANG et al., 2018; ZARDARI; JUNG, 2016; THORLEUCHTER; POEL, 2012) e clusterização (HUANG et al., 2018; GONZALEZ-COMPEAN et al., 2019; KATZ; ELOVICI; SHAPIRA, 2014). Além dos algoritmos implementados como K-NN (ZARDARI; JUNG, 2016), *Label Propagation Algorithm* (LPA) (HUANG et al., 2018), K-means (KATZ; ELOVICI; SHAPIRA, 2014) e *Logistic Regression* (THORLEUCHTER; POEL, 2012). No entanto nenhum dos estudos fez uso de redes neurais artificiais.

Dos 83 estudos resultantes da aplicação da revisão sistemática identificou-se que apenas 16 estudos avaliaram o uso de redes neurais artificiais na cibersegurança. Entretanto, as RNA apresentaram bons resultados em diferentes aplicações relacionadas à mineração de texto (WANG et al., 2016; VINODHINI; CHANDRASEKARAN, 2016). Com base nisso, foi proposto avaliar também uma implementação de RNAs como um recurso a ser empregado em soluções de cibersegurança como DLPs em frente as comumente implementações de classificação de textos.

3.2.3 Análise de Dados

Modelos de prevenção ao vazamento de dados costumam lidar com uma grande quantidade de dados. Dessa forma, visando produtividade, DLPs costumam realizar uma análise de contexto. Documentos são analisados, diferentes métricas são aplicadas e com base em um valor definido, os diversos documentos são comparados entre si a fim de classifica-los com as abordagens pertinentes.

O modelo *CoBAn* (KATZ; ELOVICI; SHAPIRA, 2014) tem por objetivo monitorar vazamentos de dados acidentais e intencionais baseados em contexto. O modelo realiza a aplicação de pré-processamento nos documentos e define um peso para cada termo. Então define uma métrica para a associação do termo com a classificação (confidencial, não confidencial) e cria outra métrica para a definição do "nível"de confidencialidade do documento.

Wan and Jin propõem uma solução para mitigar o vazamento de dados em ambientes de nuvem. O modelo extrai tópicos de cada relação do usuário a fim de estabelecer a importância da conexão com cada contato. Realiza também a criação do perfil do usuário baseado no histórico de arquivos compartilhados e nos respectivos destinatários. Por fim, calcula uma métrica baseada na conexão do usuário com cada contato (WANG; JIN, 2011).

A proposta dos autores Alneyadi, Sithirasenan and Muthukkumarasamy tem como objetivo um mecanismo para classificação de segurança das informações. A proposta define categorias, seleciona documentos, extrai *N-grams* e realiza o pré-processamento. Ao final do processo é implementada uma classificação baseada na distância dos *N-grams* (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013). Dos estudo obtidos como resultado da revisão sistemática da literatura aplicados à prevenção ao vazamento de dados, apenas 03 utilizam como abordagem de análise dos dados a análise de conteúdo (THORLEUCHTER;

POEL, 2012; DESHPANDE et al., 2015; MASHECHKIN et al., 2015).

Um dos estudos apontados é do autor Mashechkin et al., que propõem uma abordagem que tem como objetivo identificar mudanças no envio de informações corporativas com base no comportamento do usuário (MASHECHKIN et al., 2015). A proposta apresenta uma sequência definida de etapas que se inicia com a seleção de uma base de dados que possui uma ligação entre o conteúdo, usuário e tempo. Então é feita a seleção de um intervalo de tempo para realizar a modelagem de tópicos. Por fim, realiza a predição de comportamento com base no histórico do usuário (dia e tipo de informação). A abordagem do conteúdo, por ser mais profunda, fornece mais dados para uma análise semântica e então se apresenta como uma melhor abordagem na detecção da transformação de conteúdo.

3.2.4 Contexto de Aplicação

Os estudos apresentados na Tabela 3 têm como contexto de aplicação *Cloud Computing* (WANG; JIN, 2011; GONZALEZ-COMPEAN et al., 2019), documentos eletrônicos (KATZ; ELOVICI; SHAPIRA, 2014; ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013; DESHPANDE et al., 2015), correios eletrônicos (MASHECHKIN et al., 2015) e há aqueles que não definiram em seu escopo o contexto (THORLEUCHTER; POEL, 2012; ZARDARI; JUNG, 2016; HUANG et al., 2018). Os trabalhos relacionados tem como contexto específico de aplicação outros setores que não a área médica. Além disso, para seus respectivos experimentos são utilizados conjuntos de dados sintéticos (DESHPANDE et al., 2015). Cada contexto tem suas características e necessidades únicas, pois as informações estão particularmente estruturadas de forma a atender o público alvo. Dessa forma, o contexto de aplicação proposto nesse estudo, com uma base de dados real, se mostra uma oportunidade de pesquisa em sistemas de prevenção ao vazamento de dados a ser explorada.

4 MODELO ARTERIAL

Esse capítulo detalha todos os módulos que compõem o modelo ARTERIAL, um modelo de prevenção ao vazamento de informações voltado ao contexto clínico/hospitalar. O capítulo está organizado de forma a proporcionar uma melhor compreensão do modelo. O capítulo começa pela Seção 4.1 que descreve as tomadas de decisões de projeto ao longo do desenvolvimento do modelo. Em seguida, é apresentando uma visão geral do modelo. Por fim, a Seção 4.3 descreve a arquitetura proposta pelo modelo ARTERIAL.

4.1 Decisões de Projeto

O modelo ARTERIAL se propõe a tratar a informação em estado "em uso". Dessa forma, o modelo é desenhado prevendo o recebimento da informação em modo texto. Para os experimentos foram utilizados o padrão de texto XML e XMI. A implantação do modelo exige adaptação de acordo com a estrutura da organização.

O modelo tem como principal recurso de PLN o reconhecimento de entidades nomeadas (NER). Na literatura foram encontrados modelos já testados e avaliados (JIANG; BANCHS; LI, 2016; VYCHEGZHANIN; KOTELNIKOV, 2019; SCHMITT et al., 2019). No entanto, como o principal aspecto de aplicação é o idioma, o *dataset* é parte fundamental para a avaliação de desempenho da abordagem. Os modelos encontrados são aplicados e testados no idioma inglês. Afim de escolher a abordagem mais adequada ao contexto do ARTERIAL foram avaliados três modelos: LSTM, StanfordNER[1] e SpaCy[2]. O SpaCy e o StanfordNER foram selecionados por apresentarem os melhores resultados entre os trabalhos relacionados (VYCHEGZHANIN; KOTELNIKOV, 2019; SCHMITT et al., 2019).

Ainda, as RNRs têm se mostrado um caminho promissor quanto ao processamento de linguagem natural (DU; VONG; CHEN, 2021). Mesmo entre as RNR temos a LSTM como destaque. LSTM-NN faz uso de um mecanismo de atenção múltipla para ajudar a modelar relações não lineares complexas. Ainda, possui uma arquitetura bem projetada para capturar diferentes aspectos das representações de dados que correspondem a níveis distintos de abstração (TIAN et al., 2019). Desta forma, palavras e ou sentenças diferentes que levem o mesmo conteúdo podem ser identificadas e ações tomadas caso necessário. Assim, foi incluído um modelo baseado em LSTM na avaliação.

Como já mencionado, o corpus de dados é parte fundamental para o treinamento e avaliação do modelo. Os dados utilizados nesse estudo são provenientes de um projeto entre a Universidade do Vale do Rio dos Sinos (UNISINOS) e o Grupo Hospitalar Conceição (GHC). Os dados foram extraídos através de um *dump* da base de dados do hospital, dessa forma os arquivos são organizados conforme as tabelas do banco de dados. Para um melhor resultado,

[1] https://nlp.stanford.edu/software/CRF-NER.html
[2] https://spacy.io/

os dados do paciente contidos em cada tabela, foram unificados em um único arquivo.

Uma vez em posse dos dados, foram identificados 07 entidades para classificação. São as entidades: Enfermidade, Exame, Geral, Medicação, Posologia, Resultado e Sintoma. O modelo ARTERIAL prevê em um de seus módulos a validação da identificação do usuário afim de permitir ou negar o acesso a informação. Para o estudo de caso, foram considerados papeis controlados, como Geral, Enfermeiro e Médico. O papel Geral consegue ver apenas conteúdos não sensíveis, independente de sua subclassificação. Médico e Enfermeiro se diferenciam apenas no acesso a informações das entidades "Exames"e "Resultados".

4.2 Visão Geral

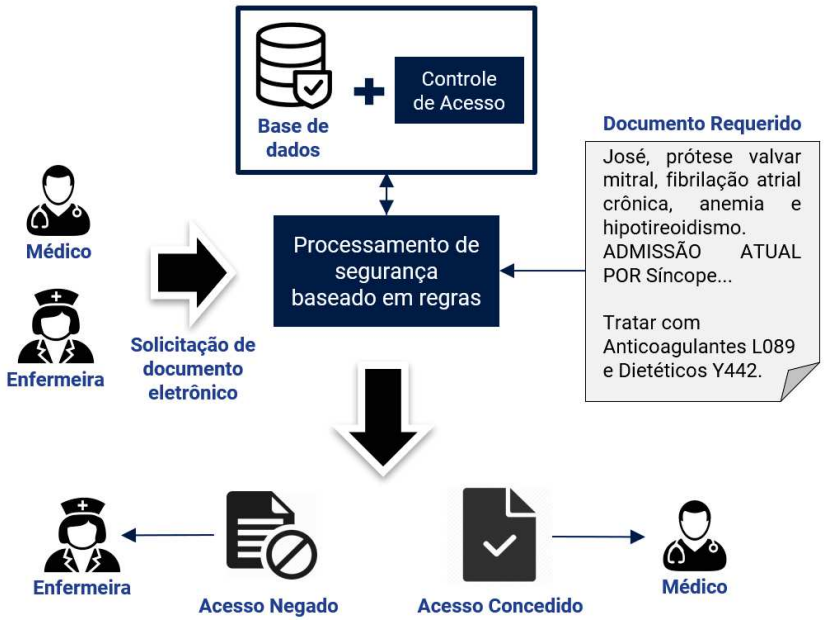
A informação em um ambiente clínico/hospitalar é um ativo relevante e volátil. A partir dessas características se observa uma necessidade imprescindível da proteção desse ativo. As informações de saúde devem ser acessíveis apenas por pessoas autorizadas (NEWAZ et al., 2020). Um dos principais aspectos do modelo ARTERIAL é tratar da confidencialidade das informações. Segundo a norma ISO/IEC:27001, que trata de requisitos para sistemas de gestão de Segurança da Informação, deve ser assegurado que a informação documentada esteja protegia adequadamente contra a perda da confidencialidade.

O modelo proposto tem por objetivo principal a prevenção ao vazamento de informações que, em resumo é tornar a informação acessível apenas a quem tenha a autorização de acessá-la. Para seu objetivo o modelo faz uso de tecnologias como a compreensão do texto escrito e a extração da informação. Para alcançarmos o objetivo do modelo proposto, encontramos no PLN (LIDDY, 2001) e nas RNAs (HAYKIN, 2009) alternativas apropriadas para compreensão de textos não estruturados e para a extração da informação.

Em geral, os modelos de DLPs são baseados em heurísticas, regras, padrões e impressões digitais (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2016). Esses modelos são inflexíveis, demandam uma alta carga de configuração e não comportam a variação linguística do idioma monitorado. Ainda, após a análise pelo DLPS, geralmente as ações de proteção tomadas consistem em notificação, auditoria, bloqueio, criptografia e quarentena. A Figura 10 exemplifica a abordagem comum de um DLP, onde após a análise do documento é realizada a tomada de decisão, quanto a permissão ou o bloqueio do acesso ao documento em sua totalidade.

O modelo ARTERIAL apresenta uma arquitetura similar ao utilizado em DLPs, se diferenciando em características como a análise do conteúdo e nas tecnologias utilizadas para a extração das informações. As informações não estruturadas são analisadas com base em seu conteúdo e significado. Diferente das abordagens atuais que utilizam expressões regulares, regras ou correspondência exata de arquivos (SECUROSIS, 2010). Somando essa análise com a implementação de uma classificação dinâmica das informações, a abordagem resultante do modelo apresentado é uma visualização parcial das informações. Como resultado, a

Figura 10 – Abordagem de acesso a informação baseado em regras



Fonte: Elaborado pelo autor.

propriedade da confidencialidade é aplicada e o compartilhamento das informações não é cerceado. Assim, é possível manter o compartilhamento de toda a informação pertinente. Onde antes uma informação poderia limitar o acesso ao documento como um todo, dessa forma, apenas a informação sensível será ocultada. A Figura 11 exemplifica a visão geral do modelo ARTERIAL.

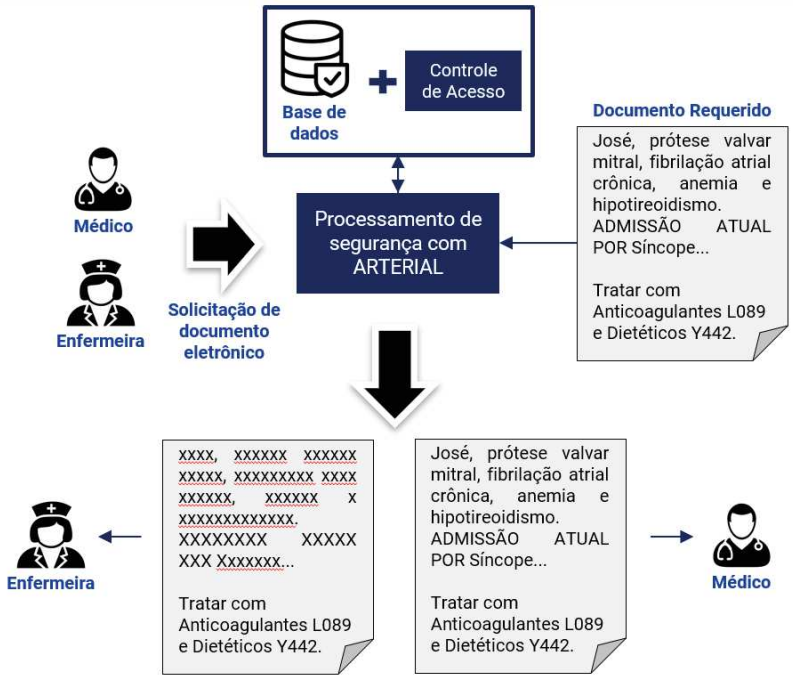
4.3 Arquitetura

Com o objetivo de promover a confidencialidade das informações o modelo foi desenvolvido e organizado em três módulos. São eles, Pré-Processamento, Extração da Informação e Validação. A Figura 12 exemplifica o esquema descrito. Cada módulo foi desenvolvido a fim de otimizar o desempenho e qualidade do modelo.

ARTERIAL é um modelo pensado para atuar em tempo de execução e funciona em conjunto com outros sistemas locais para o controle de acesso às informações. Uma vez que o modelo visa a confidencialidade das informações são necessários acessos aos documentos eletrônicos e aos papeis de identificação utilizados no domínio local. Nesse estudo os níveis e as respectivas nomenclaturas para acesso ao sistema são resumidos ao tempo "Papel". O modelo espera receber as informações sem qualquer criptografia e os papeis em modo texto.

A partir do momento que um usuário solicita a leitura de um documento, o modelo ARTERIAL é acionado e recebe então o documento e o nível de acesso do usuário. Quais papéis e quais serão as entidades acessadas são definidas pelos gestores da organização, não

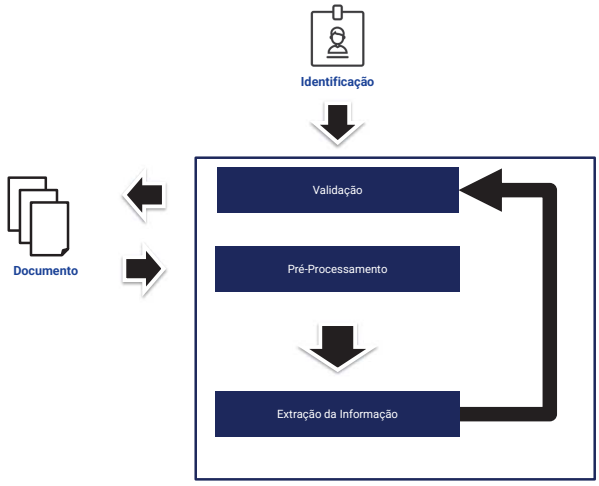
Figura 11 – Abordagem de acesso a informação baseado no modelo ARTERIAL



Fonte: Elaborado pelo autor.

sendo definidos pelo modelo. Então, o documento é pré-processado com o objetivo principal de adequar o conteúdo a estrutura necessária para a próxima etapa. Em seguida é então realizada a extração da informação, por fim a validação das informações classificadas de acordo com o nível de segurança do usuário. Nesse último módulo ainda é realizada a ocultação das entidades consideradas protegidas ao papel de acesso. Os módulos apresentados na Figura 12 são detalhados a seguir.

Figura 12 – Arquitetura modelo ARTERIAL



Fonte: Elaborado pelo autor.

4.3.1 Pré-Processamento

O autor Miner organiza o processamento de textos em três etapas principais: (a) Definição do corpus; (b) Pré-processamento; (c) Extração de conhecimento. O pré-processamento é utilizado com o objetivo de adequar dados não tratados a um formato intermediário, incrementando a extração de padrões por outras tarefas de PLN (MUNKOVÁ; MUNK; VOZÁR, 2013; ANGIANI et al., 2016; UYSAL; GUNAL, 2014).

Os autores Inzalkar e Sharma especificam três técnicas para o pré-processamento de textos: (a) Tokenização; (b) Remoção de palavras-chaves; (c) *Steamming*. Outros autores apresentam técnicas complementares como parte do discurso (PoS) que identifica e atribui marcações baseadas em gramática, como por exemplo substantivo, verbos e adjetivos (SLANKAS et al., 2014; OLIVEIRA; MERSCHMANN, 2021); e a *lemmatization* que é mais complexa e analisa a morfologia das palavras reduzindo-as a um *lemma* (SINGH; GUPTA, 2017). Geralmente são utilizados um conjunto de técnicas para preparar os dados para as tarefas de mineração de texto (FELDMAN; SANGER et al., 2007).

Para o modelo ARTERIAL foram utilizadas em conjunto as técnicas de tokenização e PoS. A tokenização tem por objetivo organizar o texto em uma sequência de termos que são definidos a partir de delimitadores e compostos por uma ou mais palavras e ou símbolos (OLIVEIRA; MERSCHMANN, 2021). Como delimitadores para os termos, podem ser utilizados espaços, pontuação ou caracteres especiais. No presente modelo, foram utilizados os espaços entre as palavras como delimitadores aos termos. O PoS realiza a marcação gramatical dos termos com base na biblioteca no idioma português-br denominado Floresta.

O contexto de aplicação deve ser levado em consideração quanto a escolha das técnicas de pré-processamento. O modelo foi desenvolvido levando em consideração dados disponíveis para treinamento e testes. Os registros utilizados são provenientes do campo evolução de prontuários eletrônicos. Dessa forma, letras maiúsculas e minusculas, assim como caracteres especiais são parte importante do conteúdo e portanto não foram desconsideradas.

4.3.2 Extração de Informação

O módulo Extração da Informação é responsável por extrair o conhecimento do texto, entendendo o significado dos termos e rotulando-o de forma a permitir o controle de acesso a seus conteúdos. Nesse módulo, se faz uso da tarefa de *text mining* denominada Extração de Informação (GUPTA; LEHAL et al., 2009; WEIGUO et al., 2005; ALLAHYARI et al., 2017).

A extração da informação tem por objetivo identificar informação estruturada em conteúdo não estruturado(AGGARWAL; ZHAI, 2012b) e tem como sub tarefas o NER e a extração de relações. Essas sub tarefas identificam entidades já conhecidas com pessoas, organizações e lugares e fazem relação entre elas (GUPTA; LEHAL et al., 2009). Inicialmente, o NER era utilizado como um sistema baseado em regras, mas estudos sugerem a utilização de métodos de

aprendizado de máquina baseados em estatísticas (AGGARWAL; ZHAI, 2012b).

O modelo proposto faz uso do StanfordNER[3] para a aplicação do reconhecimento de entidades nomeadas. A ferramenta em questão reconhece entidades levando em consideração seu treinamento. O StanfordNER tem como o idioma principal o inglês, oferecendo suporte a alguns outros como o alemão, espanhol e chines. Uma vez que o idioma português-br não tem por padrão suporte por parte da ferramenta, o presente estudo se propôs a treinar e qualificar a ferramenta para o idioma e contextos específicos. Assim, foram anotados textos ligados ao contexto clínico/hospitalar e a ferramenta foi configurada e treinada.

O modelo ARTERIAL foi desenvolvido para reconhecer um número limitado de entidades em um primeiro momento. São elas: Enfermidade, Exame, Geral, Medicação, Posologia, Resultado e Sintoma. Esse módulo recebe um texto organizado em uma sequência de termos denominamos *tokens*, realiza então a extração de informação através do NER e rotula os *tokens* com base nas entidades reconhecidas. Por fim, como saída desse módulo temos um texto rotulado onde são identificadas as entidades pertinentes ao contexto.

4.3.3 Validação

A principal função do módulo de Validação é realizar o controle de acesso ao conteúdo de acordo com o papel do usuário. O módulo tem como entrada a saída do módulo de Extração da Informação. Assim, recebe o texto rotulado com as entidades já conhecidas. Cabe então verificar os papeis cadastrados e quais são as respectivas entidades autorizadas a leitura. O modelo ARTERIAL foi desenvolvimento em um primeiro momento apenas com os papeis Geral, Médico e Enfermeiro. O papel Geral consegue ver apenas conteúdos não sensíveis, independente de sua subclassificação. Médico e Enfermeiro se diferenciam apenas no acesso a informações das entidades "Exames"e "Resultados".

Uma vez realizada a verificação do papel do usuário, com as respectivas permissões de leitura, o módulo de Validação realiza o processo de anonimização do conteúdo definido como sigiloso. É atribuído um conjunto de caracteres X a cada *token*, levando em consideração seu tamanho. Dessa forma, de todo o conhecimento extraído do texto, o usuário tem acesso apenas ao conteúdo que seu acesso permite. Assim, tem sua restrição de acesso vinculado ao conteúdo e não ao documento como um todo.

4.3.4 Considerações Finais

O modelo ARTERIAL propõem a análise de conteúdos e a extração de informações para realizar o controle de acesso a informação. Os módulos funcionam em uma sequência pré definida a fim de proporcionar o melhor desempenho de cada processo. Em um primeiro momento é realizado o pré-processamento do texto, onde o mesmo é tokenizado e preparado

[3]https://nlp.stanford.edu/software/CRF-NER.html

de forma adequada às próximas etapas. Em seguida, temos a marcação gramatical dos termos através do PoS e então a extração da informação que utiliza de técnicas como NER para identificar e rotular o significado do conteúdo. Por fim, é realizada a validação das permissões do usuário solicitante do texto com as autorizações pré definidas no modelo.

As ferramentas e abordagens aplicadas ao modelo foram selecionadas com base no resultado de experimentos realizados. Os experimentos são apresentados e discutidos na seção seguinte. Na Seção 6 são descritos os resultados e é afirmado que o modelo ARTERIAL atingiu o objetivo proposto de extrair informação de textos não estruturados e realizar uma ação, aqui posta como controlar o acesso. Ainda, na seção são apresentadas as métricas de avaliação e a eficiência do modelo.

5 MATERIAIS E MÉTODOS

Esse capitulo tem como objetivo descrever os materiais e os métodos utilizados para o treinamento e para a validação do modelo ARTERIAL. À vista disso, a Seção 5.1 descreve a base de dados, as ferramentas e as bibliotecas utilizadas ao longo da execução dos experimentos. Por fim, são descritos na Seção 5.2 os detalhes dos procedimentos realizados e das configurações utilizadas.

5.1 Descrição dos Materiais

Para a realização dos experimentos e posteriormente do modelo, foram necessários bases de dados relacionados ao tema de foco do estudo. Assim como, foram escolhidas algumas ferramentas dentre tantas presente na literatura. Essa seção se propõem a apresentar, detalhar e justificar as escolhas do presente estudo.

5.1.1 Base de Dados

O desenvolvimento do modelo proposto tem como necessário uma fonte de dados com informações médicas não estruturadas. Tal fonte de dados é necessária para o treinamento, testes e avaliação do modelo ARTERIAL. Ao longo da pesquisa para o desenvolvimento do modelo, diversas possíveis fontes de dados foram encontradas. No entanto, duas se destacaram pela relevância e número de citações; são elas: MIMIC *(Medical Information Mart for Intensive Care)*[1] e n2c2 *(National NLP Clinical Challenges)*[2].

MIMIC é um conjunto de dados aberto, desenvolvido pelo Laboratório de Fisiologia Computacional do MIT, compreendendo dados de saúde não identificados associados a aproximadamente 60.000 internações em unidades de terapia intensiva. A base de dados contempla sinais vitais, medicamentos, medições laboratoriais, observações e notas mapeadas pelos prestadores de cuidados, equilíbrio de fluidos, códigos de procedimentos, códigos de diagnóstico, relatórios de imagem, tempo de permanência no hospital, dados de sobrevivência e muito mais (JOHNSON et al., 2016).

O n2c2 foi criado originalmente em um antigo Centro Nacional de Computação Biomédica (NCBC), financiado pelo NIH, conhecido como i2b2 (*Informatics for Integrating Biology and the Bedside*). Reconhecendo o valor no texto não estruturado, o i2b2 forneceu conjuntos de notas totalmente anonimizadas para uma série de desafios e *workshops* de PLN, que posteriormente foram disponibilizadas à comunidade para fins gerais de pesquisa.

Ambas as bases de dados são relevantes, no entanto nenhuma delas possui registros no idioma português-br. Assim, usaremos uma terceira fonte de dados. A fonte de dados é

[1] https://mimic.physionet.org/

[2] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

disponibilizada através de um projeto que está sendo desenvolvido em parceria entre a Universidade do Vale do Rio dos Sinos (UNISINOS) e o Grupo Hospitalar Conceição (GHC). Os dados são advindos de prontuários e ou boletins de atendimentos. As instituições firmaram parceria através do projeto intitulado *Modelagem Semântica para a Identificação do Risco de Sepse em Sistema de Urgência e Emergência*. Foram fornecidas *dumps* de 5 (cinco) tabelas das bases de dados do GHC. O projeto que fornecerá os dados já foi submetido e aprovado pelos comitês de ética e parceiros de pesquisa da Unisinos.

Como o modelo ARTERIAL tem como foco textos não estruturados, foram utilizados os campos "Evolução"uma vez que eles são de escrita livre. Os campos "Evolução"são preenchidos por diferentes profissionais da saúde e apresentam informações relevantes. Para o estudo foram identificados e elencados alguns conteúdos, sendo eles: (a) enfermidades, (b) exames, (c) medicações, (d) posologias, (e) resultados e (f) sintomas.

5.1.2 *Natural Languagem Toolkit*

Natural Language Toolkit (NLTK) é um conjunto de módulos desenvolvidos na linguagem Python que juntos possibilitam a manipulação e processamento de dados em linguagem natural (PLN) (BIRD, 2006). A ferramenta oferece tarefas de processamento, modelos de textos anotados, ferramentas de visualização gráficas e tutoriais. Dentre os módulos estão presentes o processamento de *tokens*, atribuição de rótulos, árvores de informação e opções de estruturação de dados. Em resumo, a ferramenta possibilita a tokenização, o *steamming*, atrbuição de rótulos, classificação e clusterização. Segundo Bird (2006), o NLTK foi idealizado para a utilização de estudantes que estão aprendendo sobre NLP ou pesquisadores que conduzem pesquisas relacionadas a essa mesma temática.

5.1.3 *StanfordNER*

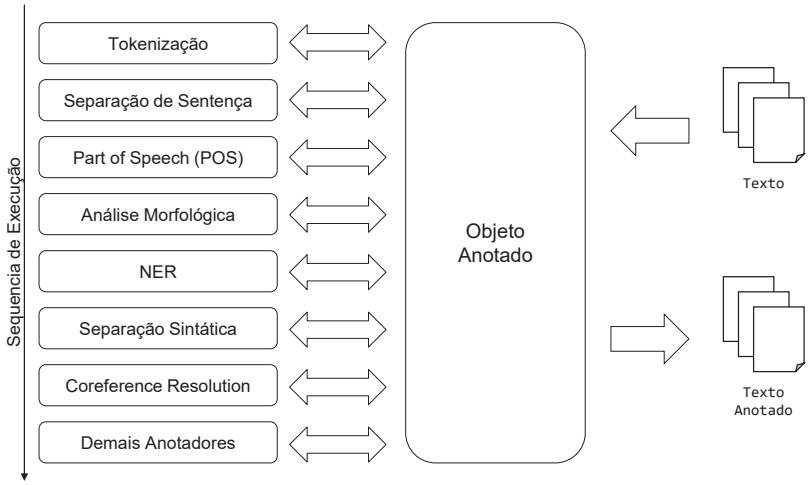
StanfordNER é um recurso pertencente ao conjunto de recursos do Stanford CoreNLP. O CoreNLP é desenvolvido na linguagem Java[3] e oferece um conjunto de recursos para o processamento da linguagem natural. Entre os recursos estão a possibilidade que os usuários derivem anotações linguísticas para texto, incluindo *token* e limites de frase, PoS, NER, entre outros.

A ferramenta Stanford CoreNLP oferece uma interface uniforme para uma entidade anotador que adiciona algum tipo de informação baseada em uma análise à algum texto. Uma entidade anotador faz isso pegando um objeto anotação ao qual pode adicionar informações extras. Essa arquitetura básica provou ser bem-sucedida e ainda é a base do sistema descrito aqui. Embora existam vários bons kits de ferramentas de análise de linguagem natural, Stanford CoreNLP é um dos mais utilizados (MANNING et al., 2014). A Figura ilustra o

[3]https://www.java.com/pt-BR/download/help/whatis_java.html

processo.

Figura 13 – Exemplificação do funcionamento do Stanford CoreNLP



Fonte: Adaptada de (MANNING et al., 2014)

StanfordNER é uma implementação Java de um NER. O NER rotula sequências de palavras em um texto que são nomes de coisas, como nomes de pessoas e empresas, ou nomes de genes e proteínas. A ferramenta possui extratores de recursos bem projetados para reconhecimento de entidade nomeada e muitas opções para definir extratores de recursos. Por definição, são disponibilizados os reconhecedores de entidades nomeadas para o inglês, especialmente para as classes Pessoa, Organização e Localização (FINKEL; GRENAGER; MANNING, 2005) .Há ainda outros modelos para diferentes idiomas e circunstâncias e é possível o treinamento de entidades e idiomas.

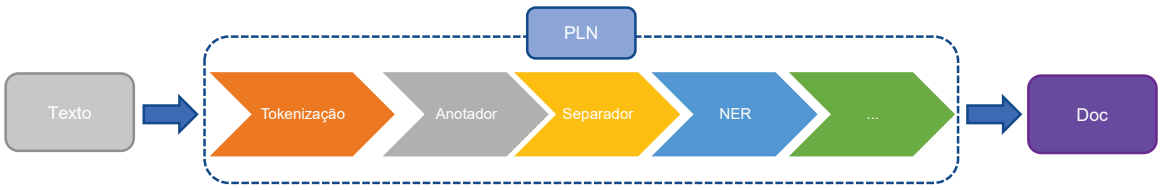
5.1.4 *SpaCy*

SpaCy é uma biblioteca para processamento avançado de linguagem natural em Python e Cython. É construído com base nas pesquisas mais recentes e foi projetado desde o primeiro dia para ser usado em produtos reais. SpaCy vem com *pipelines* pré-treinados e atualmente oferece suporte a tokenização e treinamento para mais de 60 idiomas. Ele apresenta modelos de RNA, marcação, análise, NER, classificação de texto entre outros (HONNIBAL et al., 2020). SpaCy é um software comercial de código aberto, lançado sob a licença do MIT[4].

A ferramenta tem como etapas de funcionamento: (a) primeiro transforma o texto em token para produzir um objeto Doc; (b) o Doc é então processado em várias etapas diferentes que podem incluir um *tagger*, um lematizador, um analisador e um reconhecedor de entidade. Cada componente do pipeline retorna o Doc processado, que é então passado para o próximo componente. A ferramenta tem como resultado as analises e classificações realizadas de

[4]https://www.mit.edu/

Figura 14 – Exemplificação do funcionamento do SpaCy



Fonte: Adaptada de (HONNIBAL et al., 2020)

acordo com o selecionado pelo usuário. A Figura 14 ilustra o descrito da ferramenta.

5.1.5 *CLAMP*

O Clamp (*Clinical Language Annotation, Modeling, and Processing*) é um kit de ferramentas de PNL clínico que fornece não apenas módulos de PNL de última geração, mas também um ambiente de desenvolvimento integrado e fácil de usar. Ainda, disponibiliza interface gráfica de usuário (GUI) para permitir que os usuários criem rapidamente pipelines de PNL personalizados para aplicativos individuais (SOYSAL et al., 2018).

Dentre o oferecido pela ferramenta estão soluções de treinamento através de aprendizado de máquina, customização de *pipelines* permitindo ao usuário uma melhor performance na sua utilização e a anotação e gerenciamento de *corporas*.

Clamp é implementado em Java como um aplicativo de desktop. Ele se baseia na estrutura da *Apache Unstructured Information Management Architecture (UIMA)* para maximizar sua interoperabilidade com outros sistemas baseados em UIMA. CLAMP também suporta a estrutura Apache UIMA Asynchronous Scaleout (AS) para processamento assíncrono em um ambiente distribuído (SOYSAL et al., 2018).

5.2 Experimentos do Modelo ARTERIAL

Uma vez que o modelo foi proposto e os materiais e métodos necessários ao modelo ARTERIAL foram analisados, concentramos nossos esforços nos experimentos e na implementação do modelo. A seção a seguir descreve todos os procedimentos realizados na configuração e execução dos experimentos pertinentes ao modelo proposto.

5.2.1 Anotação dos Dados

Como descrito anteriormente, o modelo proposto faz uso de uma base de dados composta por 5 (cinco) tabelas. Foi utilizada a tabela que tem como objetivo agrupar os dados de evolução dos pacientes, campo esse que é de escrita livre e não estruturada. Para o estudo

foram identificados e elencados alguns conteúdos, sendo eles: (a) enfermidades, (b) exames, (c) medicações, (d) posologias, (e) resultados e (f) sintomas.

Para a anotação dos dados foi utilizado o software Clamp[5]. Através dessa ferramenta foi possível realizar a anotação em uma interface gráfica, atribuindo os rótulos aos tokens e inclusive especificando relação entre rótulos. Foram também identificadas as relações entre os rótulos posologia e medicação e entre exame e resultados. Onde assim, o contexto da medicação, sua quantidade e aplicação foram anotadas. Assim como os resultados provenientes de seus respectivos exames.

A ferramenta apresenta duas principais versões, gratuita e licenciada. A gratuita oferece uma interface apenas em linhas de comando e a versão licenciada uma interface gráfica. A partir de um e-mail enviado à *University of Texas Health Science Center at Houston* apresentando o projeto e solicitando uma licença, a mesma foi concedida.

Os termos anotados são todos provenientes da área de saúde, no entanto os autores não tem formação direta na área de conhecimento da saúde. Dessa forma, para realizar a anotação foram utilizados recursos de consulta a fim de embasar toda leitura e reconhecimento de termos e seus significados. O material de consulta consiste principalmente na Internet e nas fontes: (a) Etimologia e Abreviaturas de Termos Médicos[6], (b) Avaliação de Enfermagem: anamnese e exame físico[7], (c) Termos Utilizados por enfermeiros em registros de evolução do paciente[8], (d) Abreviaturas Utilizadas no Hospital Getúlio Vergas[9], (e) Guia Farmacêutico do Hospital Naval Marcílio Dias[10], (f) Relação Nacional de Medicamentos Essenciais 2020[11], (g) Anotação de Enfermagem[12], (h) Expressões Médicas: glossário de dificuldades em terminologia médica[13] e (i) Glossário de Termos Médicos Técnicos e Populares[14]. O COFEN (conselho federal de enfermagem) dispõem na RESOLUÇÃO COFEN Nº 0514/2016 um guia de recomendações para registro de enfermagem no prontuário do paciente[15] que também foi utilizado.

Ao decorrer do processo, algumas definições foram realizadas. O termo "Enfermidade"foi a denominação escolhida pelos autores ao lugar de "doença". Uma vez que doença possui um significado mais restrito e patologia é o estudo de alterações estruturais, bioquímicas e funcionais das células (ROBBINS; COTRAN; KLATT, 2015). Assim, o termo Enfermidade foi utilizado como forma de unificar o sentido dos termos doença e patologia.

Devido aos dados serem provenientes de um hospital, a categoria exames se manteve única, não sendo subdivida entre ambulatorial e hospitalar. Contrações como "ampi-sulbactam"que

[5]https://clamp.uth.edu/index.php
[6]https://bit.ly/3GU6EQT
[7]https://bit.ly/3uSawiY
[8]https://bit.ly/3sSBCUG
[9]https://bit.ly/34Hjwgi
[10]https://bit.ly/3HYMQNv
[11]https://bit.ly/3sOvkVJ
[12]https://bit.ly/3rTSUBf
[13]https://bit.ly/3gSiSig
[14]https://bit.ly/33tH6wk
[15]https://bit.ly/3rSi7vR

sugerem referência a medicamentos, foram classificadas de acordo os resultados encontrados nas fontes de pesquisa já relacionadas quanto ao assunto medicina e enfermagem. Medicações, foram exclusivamente identificadas de acordo com o Guia Farmacêutico do Hospital Naval Marcílio Dias[16] e a Relação Nacional de Medicamentos Essenciais 2020[17]. O token "ampi-sulbactam"foi identificado como Ampicilina Sodica + Sulbactam Po Para Solução Injetável (2G + 1G).

Ainda, nesse primeiro momento as informações foram classificadas com apenas um rótulo. Dessa forma, por exemplo, o token "US abdome"(ultrassonografia), que foi classificado como Exame, tem seu token subsequente "pequenos calculos"classificado como Resultado. Em outro contexto a expressão "pequenos cálculos"poderia ser classificada como Enfermidade. Essa ambiguidade traz mais complexidade ao treinamento e funcionamento do modelo proposto. Então, por escolha dos autores, nesse primeiro momento apenas um rotulo foi atribuído por token. Do total de dados recebidos, para o treinamento e teste do modelo proposto foram anotadas 14988 palavras.

5.2.2 Pré-Processamento

O pré-processamento tem como objetivo principal preparar os dados para o processamento nos modelos inteligentes. Desta forma, em um primeiro momento é realizado a leitura e relacionamento dos dados com os seus respectivos rótulos de forma estruturada. Os dados são resultantes da etapa de anotação da Seção 5.2.1. Todos os arquivos com os campos de evolução são carregados e passam por uma série de passos para adequar os dados a entrada esperada para a LSTM, o StanfordNER e o SpaCy.

Para adequar os dados para o processamento na LSTM os campos de evolução são separados por sentenças. Em seguida, as sentenças são dividas em palavras. Com base nas palavras obtidas dos campos de evolução é criado um dicionário de palavras. O dicionário de palavras realiza um mapeamento da palavra para um identificador único. O identificador único será utilizado como entrada para a rede LSTM. Por fim, o mesmo processo de mapeamento para um identificador único é realizado para as tags das palavras. Um exemplo deste processo é apresentado na Figura 15. Após a tokenização das tags, foi realizado o processo de one-hot encoding. O one-hot enconding transforma cada rótulo de K classes em um vetor K-dimensional, onde a dimensão de vetores de codificação é fixado de acordo com o número de categorias. Esta codificação auxilia o processo de aprendizado das redes neurais (JIA et al., 2019)

Quanto ao StanfordNER, o processo de pre-processamento é mais simples. É necessário apenas a divisão dos campos de evolução por sentença e divisão das sentenças por palavras. Cada palavra contará com um token que identifica a que classe a palavra pertence. No StanfordNER, como a LSTM, cada palavra é processada individualmente no modelo. Por fim,

[16] https://bit.ly/3HYMQNv
[17] https://bit.ly/3sOvkVJ

Figura 15 – Exemplo do mapeamento de uma parte da sentença de um campo evolução.

	Sentence #	Word	Tag	Word_idx	Tag_idx
0	Sentence: 0	Evolução	O	439	2
1	NaN	HNSC:	O	1419	2
2	NaN	CIRURGIA	O	1312	2
3	NaN	VASCULAR#DM#	O	388	2
4	NaN	AMPUTAÇÃO	EN	1395	6

Fonte: Elaborado pelo autor.

para o processamento do campos de evolução no modelo SpaCy, as sentenças dos campos de evolução são agrupadas em uma lista. Cada elemento desta lista é composto por uma tupla, contendo a palavra e o token correspondente a palavra.

5.2.3 Extração de Informação

Como resultado do pré-processamento temos um texto organizado em uma sequência de termos. Esses termos estão rotulados em Enfermidade, Exame, Geral, Medicação, Posologia, Resultado ou Sintoma. Na etapa em questão são utilizadas três diferentes processos, uma vez que são testadas diferentes metodologias. Embora sejam testadas diferentes metodologias, todas tem como finalidade o reconhecimento das entidades elencadas acima. Como descrito, foram testadas três abordagens, duas delas seguem o modelo de ML e a terceira utiliza-se de RNA. Abaixo serão descritos os processos realizados, assim como os parâmetros, configurações e comandos utilizados ao longo dos experimentos.

5.2.3.1 LSTM

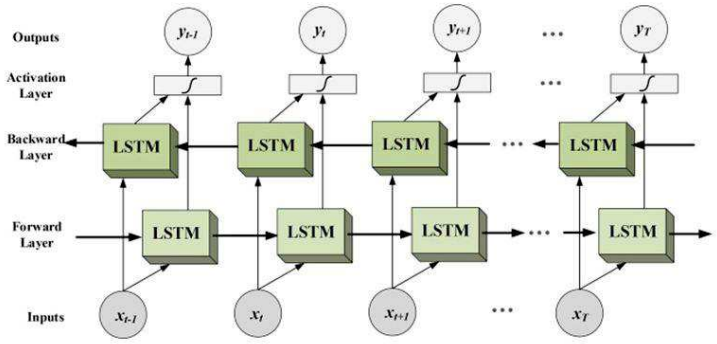
Como primeira tarefa, foi preciso planejar a arquitetura da rede neural e definir as dimensões de entrada e saída para cada camada. As RNRs, como é o caso da LSTM, são capazes de lidar com diferentes tamanhos de entrada e saída. Para esse experimento foi feito o uso da arquitetura *many to many*. Ao final do experimento a rede deve ser capaz de atribuir uma tag (y) para um token (x). A RNA faz uso de um gerador de números aleatórios para diversos fins. Visando a reprodutibilidade do experimento, definimos um valor fixo para o iniciar do gerador. Assim fica assegurado os mesmos resultados, uma vez que as demais configurações também sejam aplicadas.

Na arquitetura utilizada foram, ao todo, dispostas quatro camadas. Em um primeiro momento são dispostas três camadas, são elas: *embedding*, bi-lstm e lstm. A quarta camada é denominada *TimeDistributed Dense layer* e tem por objetivo organizar os valores de saída. A

camada de *Embedding* tem como entrada o tamanho máximo da sequência de valores. Esse tamanho é determinado a partir da quantidade de palavras presentes do conjunto de treinamento. Como resultado a camada transforma cada *token* em um vetor de N dimensões.

A segunda camada, denominada bi-lstm, consiste na aplicação de uma rede neural LSTM bidirecional. Como uma rede recorrente de duas direções, temos como resultado dois valores de dois processos que são denominados de *forward* e *backward*, Figura 16. Ainda na camada bi-lstm, é realizada a concatenação de ambos os valores resultantes dos processos *forward* e *backward* onde então é obtido o valor de saída. Nesse momento, foi preferido pela concatenação com o objetivo de apenas unir os valores, uma vez que é possível também mescla-los, multiplica-los ou realizar a média entre eles.

Figura 16 – Exemplificação do funcionamento da Bi-LSTM



Fonte: (VERMA, 2021)

Como descrito na Seção 2.3.2, uma rede LSTM funciona de forma recorrente e possui propriedades ainda mais específicas. Dentre essas propriedades está o funcionamento em blocos ao invés de camadas. Além disso, são implementados os elementos *input gate*, *forget gate*, e *output gate*. Dessa forma, a camada três é a rede LSTM propriamente dita, onde a entrada é o conjunto de dados concatenados resultantes da bi-lstm e tem como saída esperada os dados classificados em uma dimensão ainda maior.

Esse experimento teve por definição uma arquitetura de muitos para muitos (*many-to-many*) em sua RNR, de forma que se esperava uma saída para cada sequência de entrada. Podemos ter como exemplo a sequência (a1 -b1, a2 -b2 . . . an -bn), onde a e b são respectivamente a entrada e a saída de cada sequência. Como forma de concretizar o proposto, foi adicionada a camada *TimeDistributed Dense layer* que permite operações Densas (*fully-connected*) entre cada saída a cada *time-step*.

Há ainda outros elementos que são de extrema relevância na construção da extração da informação a partir do uso da LSTM. São eles: otimizador, *dropout*, quantidade de épocas e validação. Há nas redes neurais um elemento denominado função de perda que tem por objetivo guiar a rede na direção certa. Os otimizadores por sua vez atuam atualizando os parâmetros de peso para minimizar a função de perda (GOODFELLOW; BENGIO; COURVILLE, 2016). Nesse experimento foi utilizado o otimizador denominado Adam (KINGMA; BA, 2014).

Dropout é uma técnica que tem como função selecionar randomicamente neurônios da RNA durante o treinamento e então ignora-los para a etapa seguinte (GOODFELLOW; BENGIO; COURVILLE, 2016). Essa técnica tem como entrada um valor a ser definido antes do inicio da rede. Cada RNA necessita de um valor e cada valor proporciona um impacto diferente. Uma vez que a arquitetura proposta nesse subseção implementa duas RNAs dentre suas camadas, temos um valor de *dropout* para cada camada e um valor de descarte entre essas duas camadas. Para a rede biLSTM foi definido o valor de *dropout* de 0.2, já para a LSTM foi definido o valor de 0.5 e entre as camadas os valores foram de 0.2 e 0.5 respectivamente.

São definidos também os parâmetros de épocas. As épocas indicam a quantidade de vezes que o conjunto de dados de treinamento passou pelo algoritmo de aprendizado. Foi definido um valor de 100 épocas. Por fim, temos a validação da rede que foi difinida como 20% do conjunto de dados. Diversos valores foram testados nos elementos descritos, foram selecionados os que apresentaram os melhores resultados.

5.2.3.2 SpaCy

O SpaCy é uma ferramenta projetada para a aplicação em produtos reais. Dessa forma, a ferramenta oferece diversas funcionalidades pré-treinadas e disponíveis para diversos idiomas incluindo o português. Para o experimento em questão foi utilizado o módulo NER disponibilizado pela ferramenta, ajustando as configurações, aplicando o idioma português e treinando o SpaCy ao conjunto de dados e entidades descritas em seções anteriores.

Inicia-se com a instalação de pacotes necessários ao funcionamento da ferramenta e do módulo. Como primeiro passo foi instalado o pacote "conda"da linguagem Python. Em seguida, no site da ferramenta, é escolhido o pacote de linguagem. Como mencionado anteriormente, foi selecionado o pacote do idioma português. Esse pacote é denominado "pt_core_news_sm"e contém dentre seus componentes um *tok2vec*, um *parser*, um *lemmatizer*, um *morphologizer* e um *ner*.

A próxima etapa consiste na organização dos dados no formato utilizado pela ferramenta. Essa organização é realizada na etapa de pré-processamento já descrita na Seção 5.2.2. Os dados são recebidos na etapa de extração da informação no formato exemplificado na Figura 17. Foram então definidas algumas variáveis como número de iterações, equivalentes as épocas da LSTM, e o caminho de diretório para salvar o novo modelo.

Para o experimento em questão foi tomado como prioridade o treinamento do módulo de NER. Assim, apenas o NER foi mantido ativo e os demais *pipelines* desabilitados. O treinamento foi realizado então através do comando "nlp_update", com os parâmetros de *dropout*, otimizador, texto e rótulos. Para otimizar os pesos do treinamento do SpaCy foi utilizado o algoritmo *Stochastic Gradient Descent* (SGD) (BOTTOU, 2012). Por fim, a ferramenta realiza o treinamento levando em consideração os parâmetros aplicados e a quantidade de épocas definidas.

Figura 17 – Formato dos dados SpaCy

```
TRAIN_DATA = [
     ('Who is Nishanth?', {
         'entities': [(7, 15, 'PERSON')]
     }),
      ('Who is Kamal Khumar?', {
         'entities': [(7, 19, 'PERSON')]
     }),
     ('I like London and Berlin.', {
         'entities': [(7, 13, 'LOC'), (18, 24, 'LOC')]
     })
]
```

Fonte: Elaborado pelo autor.

5.2.3.3 StanfordNER

Como forma de implementar o stanfordNER na extração de informação foi utilizado o pacote do NLTK. Como já descrito anteriormente, o NLTK é um conjunto de módulos desenvolvidos na linguagem Python que juntos possibilitam a manipulação e o processamento de dados em PLN.

Uma vez instalado o pacote do NLTK, é então realizada a validação da versão e o caminho do java instalado. Para o experimento em questão é utilizado a versão 8 do java. A próxima etapa do experimento contou com a organização dos dados em duas colunas, sendo elas o token e classificação do token respectivamente, Figura 18.

Figura 18 – Formato dos dados StanfordNER

```
En O
2017 DATE
, O
Une O
intelligence O
artificielle O
est O
en O
mesure O
de O
développer O
par O
elle-même O
Super PERSON
Mario PERSON
Bros PERSON
. O
```

Fonte: Elaborado pelo autor.

Uma vez os dados pré-processados e organizados, são então definidas as variáveis de

Tabela 4 – Parâmetros aplicados ao experimento do StanfordNER

Parâmetro	Valor
useClassFeature	true
useWord	true
useNGrams	true
noMidNGrams	true
maxNGramLeng	6
usePrev	true
useNext	true
useSequences	true
usePrevSequences	true
maxLeft	1
useTypeSeqs	true
useTypeSeqs2	true
useTypeySequences	true
wordShape	chris2useLC
useDisjunctive	true

Fonte: Elaborado pelo autor.

configuração do modelo através de um arquivo de texto passado como parâmetro ao treinamento. Como destaque temos as definições de "map"que define qual coluna é entendida como token e qual é entendida como anotação.

Além disso, foi configurado também o uso de Ngrams através da variável "useNGrams"e a definição do tamanho máximo desse Ngram, definido como 6 nesse experimento, pela variável "maxNGramLeng". Outro parâmetro a se destacar nas configurações é o "wordShape"que junto de outros define o formato das palavras. Para esse parâmetro foi definido a função "chris2useLC"como formato padrão. As variáveis e seus respectivos valores são apresentados na Tabela 4

O treinamento é realizado através de um comando via terminal, o mesmo é apresentado na Figura 19. Ao comando foi passado o arquivo contendo os parâmetros já discutidos, a biblioteca e o *pipeline* a ser utilizado. Ainda, se destaca o valor -mx4g"onde esse se refere ao quantitativo de memória destinada ao treinamento, sendo "mx"a *flag* e "4g"o valor de 4 gigabytes de memória. Como resultado desse processo o modelo a ser utilizado é gerado. Para utiliza-lo passamos o modelo gerado e a *engine* do StanfordNER junto ao texto a ser reconhecido.

Figura 19 – Comandos para a execução do treinamendo do StanfordNER

```
java -cp "stanford-ner.jar:lib/*" -mx4g
edu.stanford.nlp.ie.crf.CRFClassifier -prop train/prop.txt
```

Fonte: O Autor.

5.3 Validação Qualitativa

Para avaliar qualitativamente os resultados de cada modelo, foi proposto um algoritmo para a substituição das entidades identificadas como sensíveis baseadas no papel do usuário que está utilizando o sistema de momento. Aqui cabe ressaltar, que o modelo ARTERIAL, em teoria, é capaz de suportar infinitas papeis de usuário, estando limitado apenas a capacidade computacional do servidor da instituição hospitalar. O Algoritmo 1 apresenta a lógica que realiza a supressão das entidades não permitidas para o usuário atual. Nela, as entidades identificas como não autorizadas tem suas letras substituídas pela letra X. A letra X foi escolhida arbitrariamente e apenas para fins de representação neste trabalho, podendo ser substituída por qualquer outro caractere que a instituição de saúde julgar pertinente.

Algorithm 1 Algoritmo de censura das informações sensíveis.

Require: $papel\_usuario$, $allowed\_tags$
Ensure:
  for $word, tag$ in $processed\_text$ **do**
    if tag is not in $allowed\_tags$ **then**
      $word \leftarrow 'X'$ * len($word$)
    end if
  end for

5.4 Métricas de Avaliação

A avaliação do desempenho do modelo é baseada em métricas derivadas da matriz de confusão. Em problemas de classificação binária, por exemplo, os dados são mapeados para duas classes: positiva ou negativa. Dessa forma, em um classificador qualquer é possível relacionar as previsões e os rótulos das instâncias e formar quatro situações para a classificação: Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN). Os VP são a soma de todas as previsões positivas corretas. Os FP são todas as previsões positivas incorretas. Enquanto isso, os VN são todas as previsões negativas corretas. E, por fim, as FN são todas as previsões negativas incorretas. A Tabela 5 apresenta de forma visual a organização de uma matriz de confusão genérica. A análise individual destes quatro indicadores já pode direcionar um entendimento do desempenho do modelo. Contudo, uma análise mais eficiente do modelo pode ser realizada avaliando métricas de desempenho derivadas desta matriz de confusão.

Em certas etapas do modelo proposto, o problema a ser resolvido envolve multi-classes. Assim, é necessário generalizar a matriz de confusão para utilizar essas métricas para a avaliação do modelo. Formalmente, para um conjunto de dados com classes C_k, um array de tamanho $k \times k$ é definido. Cada célula $[i, j]$ representa a frequência de observação da classe real C_i e da classe inferida C_j. Esta matriz pode ser representada em até k matrizes de

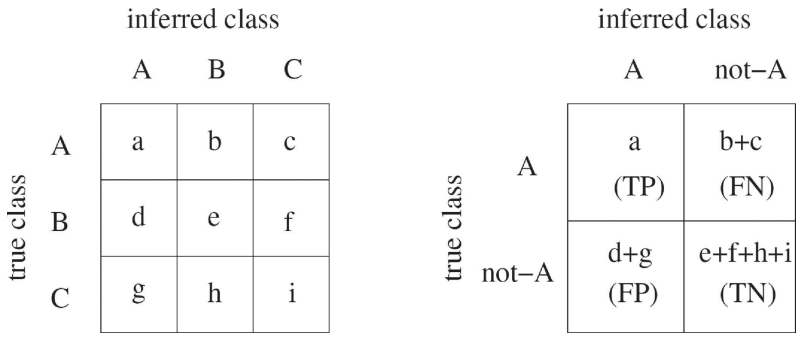
Tabela 5 – Matriz de Confusão

		Rótulo	
		Positivo	*Negativo*
Previsto	*Positivo*	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: Adaptado de Fawcett (2006)

confusão binárias, uma para cada classe C_i (RUUSKA et al., 2018). Uma representação dessa transformação pode ser vista na Figura 20.

Figura 20 – Um exemplo de uma matriz de confusão 3×3 para as classes A – C (esquerda) e a matriz de confusão binária correspondente para a classe A (direita).



Fonte: Ruuska et al. (2018).

Desta forma, e com base nos trabalhos relacionados, a avaliação dos modelos será realizada utilizando as seguintes métricas: Acurácia, Precisão e Recall. Outra maneira de avaliar o desempenho dos modelos, neste caso de forma gráfica, é pela curva ROC. Nas próximas subseções é explorada em detalhes cada uma das métricas de desempenho.

5.4.1 Acurácia

A acurácia mensura a capacidade do modelo de classificar corretamente cada caso (RUUSKA et al., 2018), e é definida como:

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \tag{5.1}$$

5.4.2 Precisão

A precisão indica o percentual verdadeiramente positivo entre todas as classificações positivas, e é definida como:

$$Precisao = \frac{VP}{VP + FP} \tag{5.2}$$

5.4.3 Recall

O *Recall*, ou sensibilidade, é a capacidade do modelo identificar a presença dos casos positivos, sendo definido como:

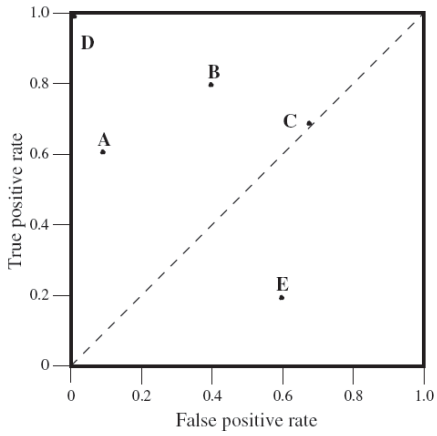
$$Recall = \frac{VP}{VP + FN} \tag{5.3}$$

5.4.4 ROC

A curva ROC é uma forma de avaliação visual de um modelo classificador. A curva ROC é plotada em um gráfico 2D, onde a taxa de verdadeiros positivos é plotada no eixo Y e a taxa de falsos positivos é plotada no eixo X. As taxas são calculadas utilizando diferentes limites de decisões (CARTER et al., 2016). Para a análise do desempenho a partir da curva ROC é necessário uma atenção especial para 3 pontos do gráfico. O ponto $(0, 0)$ indica um modelo que classifica todas as instâncias como negativas. Já o ponto $(1, 1)$, indica um modelo que classifica todos os casos como positivos. Por fim, o ponto $(0, 1)$ indica uma estratégia perfeita, onde o modelo é capaz de classificar corretamente todas as amostras do conjunto de dados (RUUSKA et al., 2018).

Dessa maneira, o objetivo na avaliação de um modelo é que o resultado fique o mais próximo possível do ponto $(1, 0)$. Além disso, a partir do cálculo da área abaixo da curva ROC, podemos obter outra métrica de desempenho a *Area Under ROC Curve* (AUC). A AUC pode ser utilizada para mensurar o quão bem um modelo pode diferenciar classes (PEDRO; MACHADO-LIMA; NUNES, 2019). Além disso, quanto maior o valor de AUC, melhor será o desempenho do modelo (RUUSKA et al., 2018). Na Figura 21 é apresentado um exemplo de uma curva ROC.

Figura 21 – Exemplo de curva ROC com cinco classificadores discretos.



Fonte: Ahmad e Yusoff (2013).

6 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados e as discussões dos experimentos realizados para validar o modelo ARTERIAL. A metodologia utilizada para realizar os experimentos é descrita no Capítulo 5. Os experimentos descritos neste capítulo foram realizados no Google Colaboratory[1]. O capítulo está dividido em duas seções. A Seção 6.1 apresenta o desempenho dos modelos StanfordNER, Spacy e LSTM. Em seguida, na Seção 6.3, apresentamos uma comparação dos resultados obtidos com o modelo ARTERIAL em relação a literatura de referência identificada e descrita no Capítulo 3.

6.1 Treinamento dos Modelos

Nesta seção, são apresentados o desempenho e os resultados para os modelos StanfordNER, Spacy e LSTM. As classes analisadas nos experimentos são: Sintomas (ST), Resultados (RS), Posologia (PS), Entidade geral (NN), Medicamento (MD), Exame (EX) e Enfermidade (EN). Com base nestas classes, foram definidas as métricas de avaliação (Seção 5.4 do modelo ARTERIAL. Portanto, nas próximas subseções, são apresentados de forma quantitativa e qualitativa os resultados obtidos para cada um dos modelos avaliados neste estudo.

6.1.1 StanfordNER

O StanfordNER oferece um conjunto de modelos pré-treinados para o reconhecimento de entidades. Além disso, o StanfordNER permite realizar customizações nas entidades e o *fine tunning* de modelos de reconhecimento de entidades já estabelecidos. Neste contexto, foi utilizado um modelo de reconhecimento de entidades treinados para um corpus da língua portuguesa e realizado o *fine tunning* para o corpus de prontuários eletrônicos apresentado na Seção 5.2.1.

6.1.1.1 Treinamento

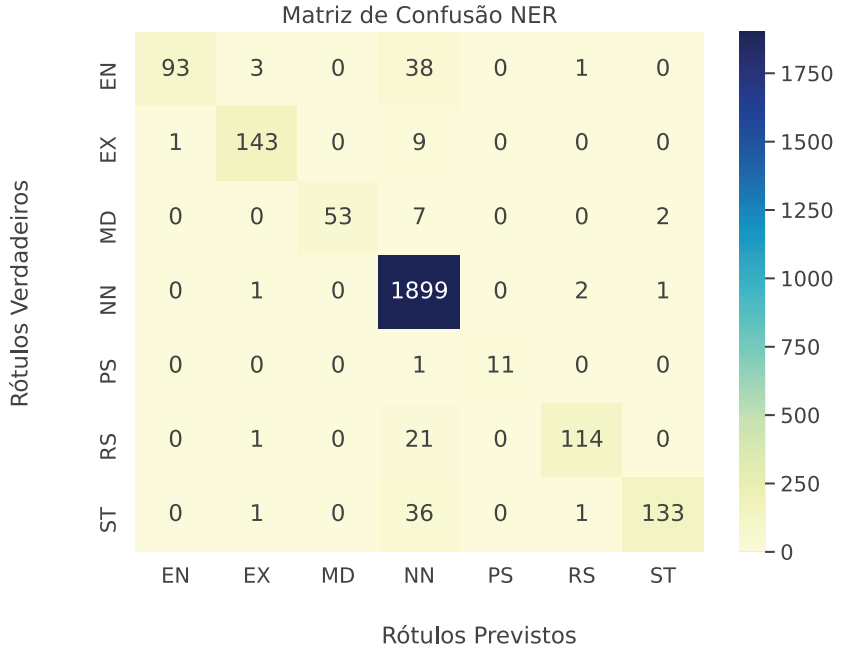
O treinamento do modelo StanfordNER foi realizado em um notebook com processador Intel(R) Core(TM) i7-10510U CPU@ 1.80GHz - 2.30 GHz, com 16GB de memória. A configuração necessária utilizada para o treinamento foi a padrão, fornecida pelo próprio StanfordNER. Como entrada de dados, utilizamos o formato do padrão do StanfordNER, seguindo a metodologia definida na Seção 5.1.3. Como resultado do processo de treinamento, é gerado um modelo treinado para o nosso conjunto de treinamento. Nas próximas seções, apresentamos os resultados para o conjunto de teste do nosso conjunto de dados.

[1] https://colab.research.google.com/?utm$_source$=scs-index

6.1.1.2 Avaliação Quantitativa

Os resultados para a avaliação quantitativa para o modelo StanfordNER são baseados no conjunto de teste. Neste sentido, foi gerada a matriz de confusão (Figura 22) para as entidades utilizadas neste trabalho. Com base nos valores obtidos na matriz de confusão, é possível gerar as métricas de Precisão, Recall e F1-Score. Na Tabela 6 são apresentadas as métricas de desempenho para o StanfordNER. Por fim, é possível avaliar visualmente os resultados a partir da curva ROC. Neste sentido, na Figura 23 é apresentada a curva ROC obtida para o conjunto de teste do StanfordNER.

Figura 22 – Matriz de confusão resultante do treinamento do StanfordNER.



Fonte: Elaborado pelo autor.

Analisando a matriz de confusão disponível na Figura 22, é possível observar que o modelo StanfordNER foi capaz de classificar corretamente a maioria das entidades do conjunto de teste. Cabe destacar, que apesar de um desempenho satisfatório, indicando uma aprendizagem da estrutura de escrita dos prontuários dos pacientes, o modelo apresentou dificuldades para distinguir alguns casos de enfermidades, sintomas e resultados de exames confundindo estes com entidades normais.

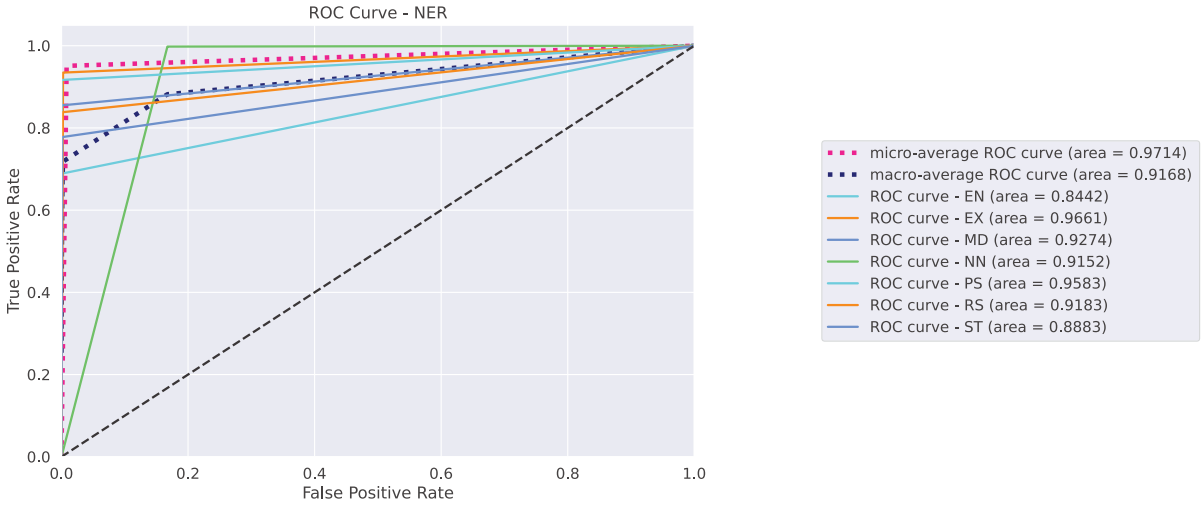
Os resultados para as métricas de desempenho apresentadas na Tabela 6 resumem, novamente, o desempenho acurado do modelo StanfordNER para o reconhecimento das entidades do conjunto de teste. O modelo obteve uma acurácia geral de 95%. Analisando isoladamente o desempenho das métricas de desempenho a termos de precisão, recall ou F1-Score, podemos perceber uma queda considerável no desempenho para as entidades enfermidade e sintomas se comparadas aos demais grupos.

Tabela 6 – Resultados para as métricas de avaliação no conjunto de testes para o modelo StanfordNER.

Classe	Precisão	Recall	F1-Score
PS	100.0	92.0	96.0
ST	98.0	78.0	87.0
MD	100.0	85.0	92.0
EX	96.0	93.0	95.0
RS	97.0	84.0	90.0
EN	99.0	69.0	81.0
NN	94.0	100.0	97.0
Macro Avg	98.0	86.0	91.0
Weighted Avg	95.0	95.0	95.0
Acurácia geral			95.0

Fonte: Elaborado pelo autor.

Figura 23 – Curva ROC resultante do treinamento do StanfordNER.



Fonte: Elaborado pelo autor.

Por fim, a área abaixo da curva ROC geral (Figura 23), do inglês *area under the ROC curve (AUC)*, é uma forma de medir quantitativamente o desempenho do modelo em termos da curva ROC. Neste sentido, quanto mais próximo de 1 melhor o desempenho do modelo. É possível mensurar essa medida de duas formas: macro e micro. A macro não considera o desbalanceamento dos dados, por outro lado a micro calcula uma medida proporcional entre as classes, eliminando assim o viés do desbalanceamento dos dados (ZEISER et al., 2021b). Com base nos valores obtidos para o StanfordNER, é possível perceber que o pior desempenho foi de 0.8442 para a entidade enfermidade. Já o melhor desempenho foi obtido para a entidade exame com 0.9661. De maneira geral, a macro AUC foi de 0.9168, o que indica que o modelo StanfordNER foi capaz de reconhecer a maioria das entidades do conjunto de teste.

6.1.2 Spacy

O SpaCy, como já mencionado, é um *framework* com um conjunto de métodos de PNL. Neste contexto, essa subseção analisa os resultados obtidos para o modelo conforme os parâmetros definidos na Seção 5.2.3.2.

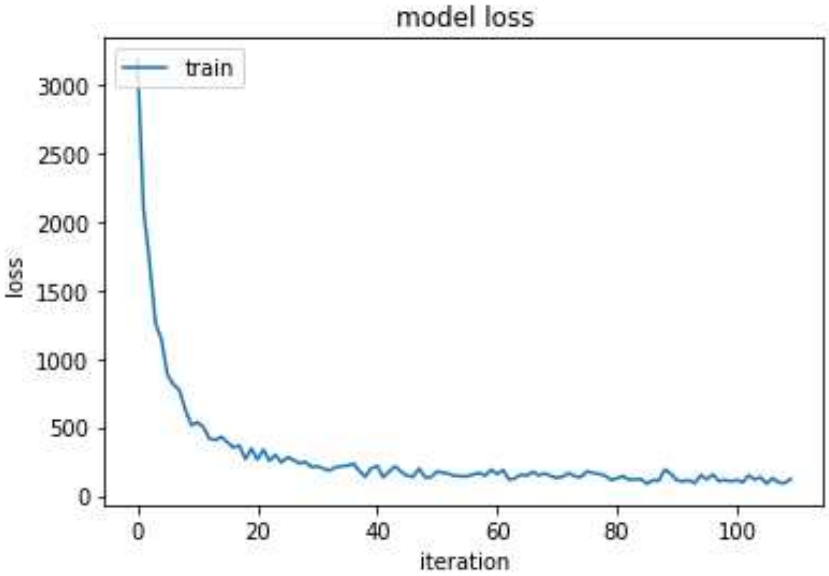
6.1.2.1 Treinamento

Para o Spacy foi utilizada a biblioteca padrão disponível no Python 3.7.11. Foram realizados quatro testes utilizando as seguintes quantidades, interações e *dropout*. São eles:

- 110 interações e dropout 0.2;
- 110 interações e dropout 0.4;
- 400 interações e dropout 0.2;
- 400 interações e dropout 0.4;

O otimizador utilizado durante o processo de *backpropagation* foi o StocaticGradientDescent. Nas Figuras 24, 25, 26 e 27 são apresentadas as curvas de loss durante o processo de treinamento para cada um dos modelos.

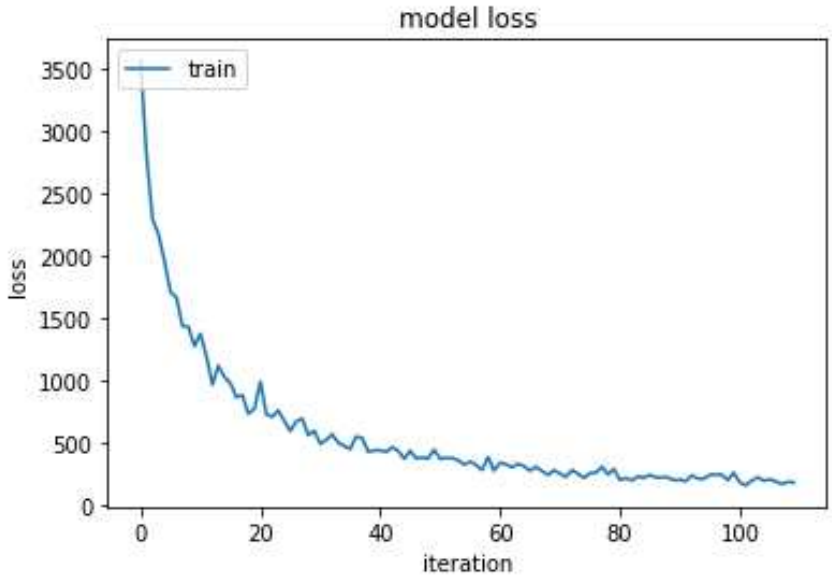
Figura 24 – Perdas resultantes do treinamento do SpaCy. 110 + 0.2



Fonte: Elaborado pelo autor.

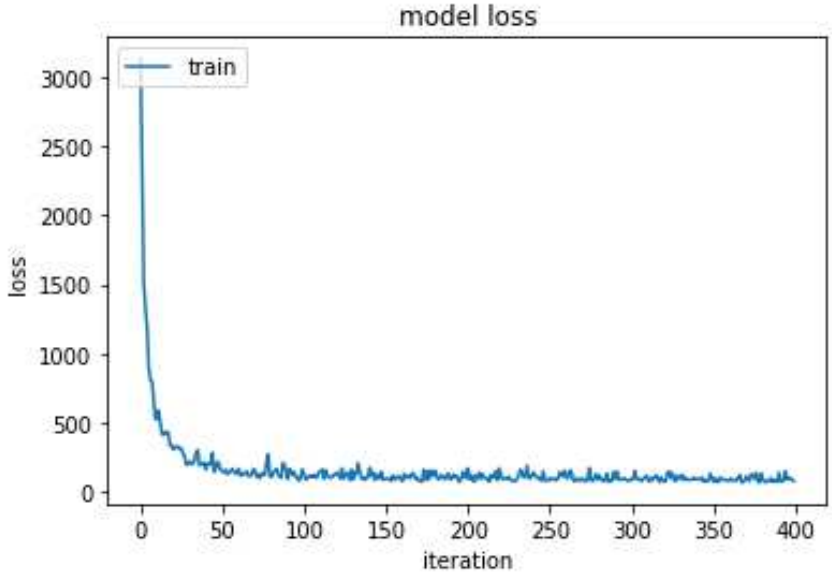
Pela curva de loss das Figuras 24, 25, 26 e 27, é possível perceber que não houve overfitting em nenhum dos quatro modelos avaliados. Além disso, a estabilização do processo de aprendizagem para os modelos ocorre entre as épocas 20 e 40 indicando que os modelos

Figura 25 – Perdas resultantes do treinamento do SpaCy. 110 + 0.4



Fonte: Elaborado pelo autor.

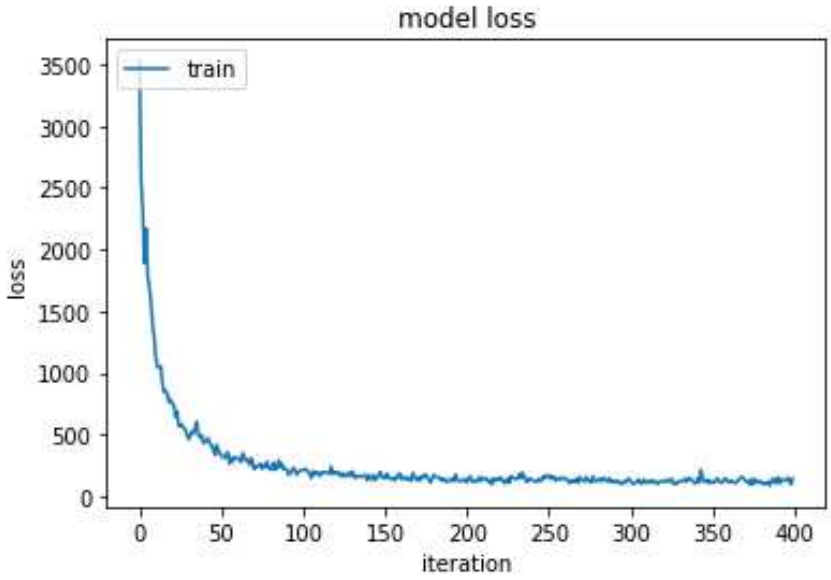
Figura 26 – Perdas resultantes do treinamento do SpaCy. 400 + 0.2



Fonte: Elaborado pelo autor.

atingiram a convergência do processo de otimização dos pesos utilizando o gradiente descendente. Os modelos com um dropout menor atingiram a convergência de maneira mais acelerada. Este comportamento pode estar relacionado com a própria definição de dropout que desliga aleatoriamente uma quantidade predeterminada de neurônios dos modelos para garantir que ocorra o aprendizado de uma determinada característica por diferentes grupos de neurônios.

Figura 27 – Perdas resultantes do treinamento do SpaCy. 400 + 0.4



Fonte: Elaborado pelo autor.

6.1.2.2 Avaliação Quantitativa

Nas Tabelas 7, 8, 9 e 10 são apresentados os resultados para cada um dos testes para o Spacy. Dada as características da biblioteca Spacy não foi possível gerar a curva ROC e a matriz de confusão.

Tabela 7 – Resultados para as métricas de avaliação no conjunto de testes para o modelo Spacy. 110 + 0.2

Classe	Precisão	Recall	F1-Score
PS	88.89	66.67	76.19
ST	92.92	81.4	86.78
MD	85.0	85.0	85.0
EX	80.34	68.12	73.73
RS	74.07	64.52	68.97
EN	62.9	84.78	72.22
Macro Avg	80.69	75.08	77.15

Fonte: Elaborado pelo autor.

De maneira geral, as métricas de desempenho para os modelos Spacy (Tabelas 7, 8, 9 e 10) variaram entre 73.69% a 87.27%. Neste sentido, é importante ressaltar a dificuldade dos modelos identificarem a posologia (PS) em todos os testes. Além disso, houve uma dificuldade para todos os modelos, exceto o de 400 interações com 0.4 de dropout, no reconhecimento das entidades exame (EX) e resultado (RS). O resultado (RS) pode ser confundido em determinadas situações devido as características numéricas dos resultados de alguns exames. Por fim, cabe destacar que o melhor desempenho para o Spacy foi obtido para 400 interações e um dropout

Tabela 8 – Resultados para as métricas de avaliação no conjunto de testes para o modelo Spacy. 110 + 0.4

Classe	Precisão	Recall	F1-Score
PS	88.89	66.67	76.19
ST	87.29	79.84	83.4
MD	85.37	87.5	86.42
EX	85.19	66.67	74.8
RS	75.0	67.74	71.19
EN	63.49	86.96	73.39
Macro Avg	80.87	75.9	77.56

Fonte: Elaborado pelo autor.

Tabela 9 – Resultados para as métricas de avaliação no conjunto de testes para o modelo Spacy. 400 + 0.2

Classe	Precisão	Recall	F1-Score
PS	80.0	66.67	72.73
ST	90.18	78.29	83.82
MD	78.05	80.0	79.01
EX	75.63	65.22	70.04
RS	76.79	69.35	72.88
EN	70.37	82.61	76.0
Macro Avg	78.5	73.69	75.75

Fonte: Elaborado pelo autor.

de 0.4.

6.1.3 LSTM

Nesta seção, são apresentados os resultados para o teste utilizando um modelo LSTM. A LSTM é definida como uma rede neural recorrente com capacidade de memória de longo prazo (CHOLLET, 2018). Nas próximas subseções são apresentados os resultados para o treinamento, avalição quantitativa e qualitativa.

6.1.3.1 Treinamento

O treinamento do modelo LSTM foi realizado utilizando a biblioteca Keras com o backend do TensorFlow. Como entrada do modelo, as palavras das sequências foram transformadas em identificadores únicos e a cada palavra foi associada um vetor no formato de one-hot encoding. O tamanho máximo de sequência considerada foi de 477 palavras. Por fim, o modelo foi treinado por 100 épocas utilizando como função de loss a categorical crossentropy, Adam para otimizador de pesos e um learning rate de 0.0005. Nas Figuras 28 e 29 apresentamos,

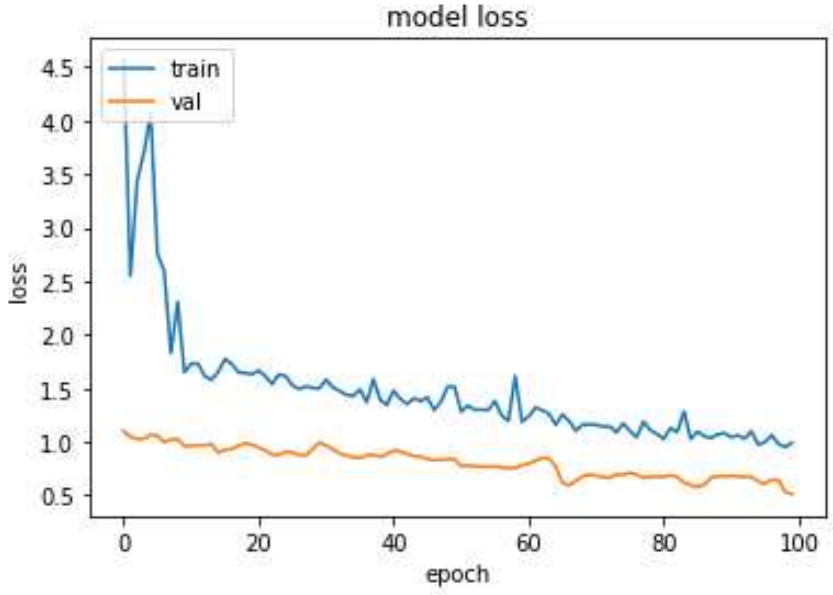
Tabela 10 – Resultados para as métricas de avaliação no conjunto de testes para o modelo Spacy. 400 + 0.4

Classe	Precisão	Recall	F1-Score
PS	100.0	66.67	80.0
ST	87.72	77.52	82.3
MD	89.74	87.5	88.61
EX	85.09	70.29	76.98
RS	84.62	70.97	77.19
EN	76.47	84.78	80.41
Macro Avg	87.27	76.29	80.92

Fonte: Elaborado pelo autor.

respectivamente, o desempenho da LSTM quanto a loss e acurácia durante o treinamento.

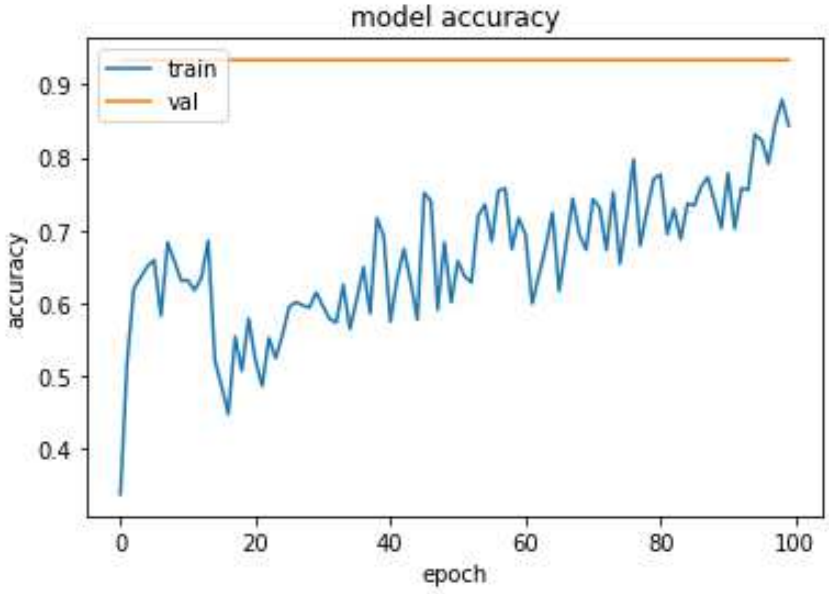
Figura 28 – Perdas resultantes do treinamento da LSTM.



Fonte: Elaborado pelo autor.

Na Figura 28, é possível identificar oscilações durante o processo de treinamento. Estas oscilações, provavelmente estão, novamente, relacionadas com o processo de dropout do modelo. Oscilações no processo de aprendizagem são comuns quando se utiliza o dropout (ZEISER et al., 2021a). Esta característica de oscilação pode indicar que o conjunto de neurônios utilizado em determinada época não foi capaz de classificar corretamente as entidades do conjunto de treinamento e validação. Contudo, ao final do treinamento o modelo atingiu uma estabilidade para o conjunto de treinamento e validação indicando um convergência do modelo.

Figura 29 – Acurácia resultante do treinamento da LSTM.



Fonte: Elaborado pelo autor.

6.1.3.2 Avaliação Quantitativa

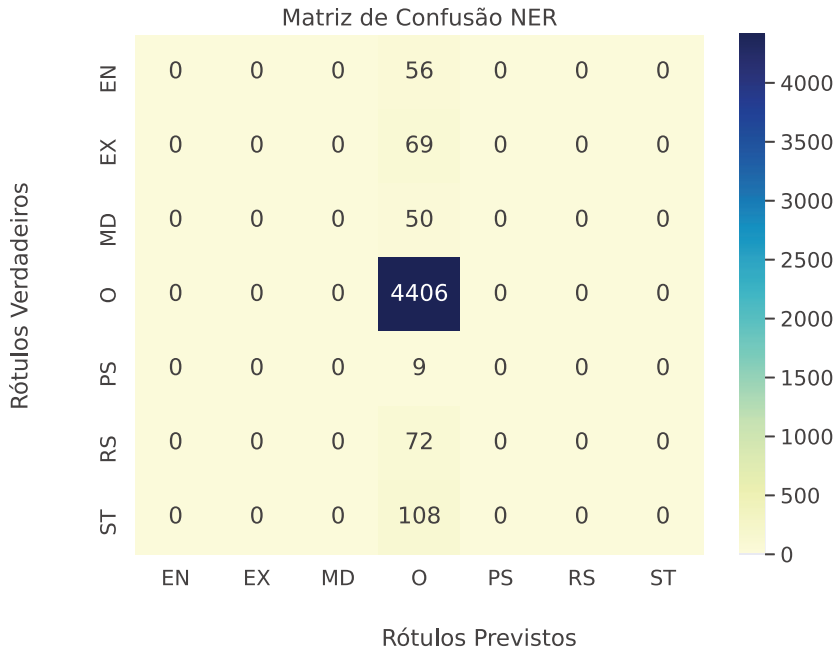
É importante mensurar o desempenho de uma arquitetura neural com base em métricas de avaliação em um conjunto de dados não utilizado durante o treinamento, aferindo assim a generalização do modelo. O conjunto de dados para o modelo LSTM foi dividido em 72% para treinamento, 18% para validação e 10% para teste. Os pesos utilizados para a avaliação do modelo LSTM se referem à época com a menor *loss* no conjunto de validação. Na Figura 30 são apresentados os valores para a matriz de confusão utilizando o modelo LSTM. Com base nestes valores são geradas as métricas de desempenho da Tabela 11.

Tabela 11 – Resultados para as métricas de avaliação no conjunto de testes para o modelo LSTM.

Classe	Precisão	Recall	F1-Score
PS	00.0	00.0	00.0
ST	00.0	00.0	00.0
MD	00.0	00.0	00.0
EX	00.0	00.0	00.0
RS	00.0	00.0	00.0
EN	00.0	00.0	00.0
0	92.0	100.0	96.0
Macro Avg	13.0	14.0	14.0
Weighted Avg	85.0	92.0	89.0
Acurácia geral			92.0

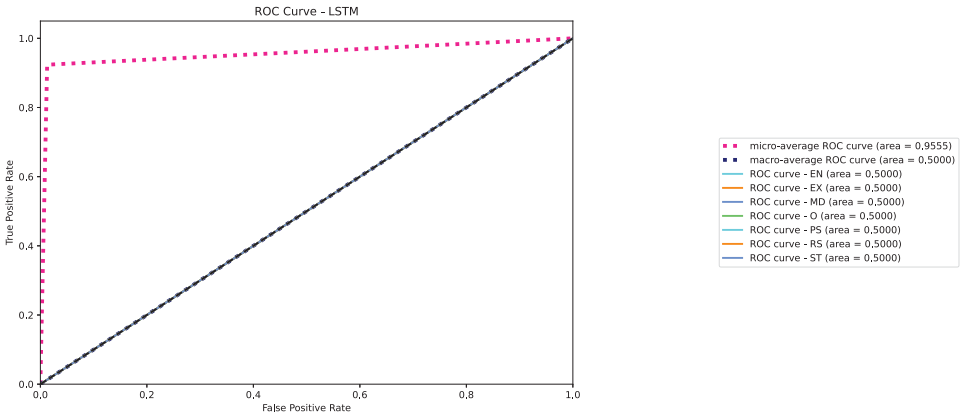
Fonte: Elaborado pelo autor.

Figura 30 – Matriz de confusão da LSTM.



Fonte: Elaborado pelo autor.

Figura 31 – Curva ROC da LSTM.



Fonte: Elaborado pelo autor.

Diferentemente dos modelos StanfordNER e Spacy, o modelo LSTM apresentou um desempenho insatisfatório para a detecção das entidades no contexto dos prontuários eletrônicos de pacientes. O resultado para a LSTM pode indicar que não houve um conjunto de dados significativo o suficiente para permitir a generalização do modelo. Este processo não ocorreu com os modelos StanfordNER e Spacy pela estrutura de refinamento. Os modelos são ajustados partindo de um modelo já pré-treinado para a língua portuguesa.

6.2 Avaliação Qualitativa

Nesta seção é apresentada uma avaliação qualitativa para o processo de reconhecimento de entidades para o melhor resultado quantitativo. Desta forma, nas Figuras 32 e 33 são apresentadas amostras do conjunto de sequências de teste para o modelo StanfordNER. A Figura 32 apresenta o campo evolução sem nenhum tratamento. Na Figura 33, as letras **X** representam palavras ou numerais suprimidos pelo modelo StanfordNER.

Figura 32 – Exemplo de campos de Evolução do conjunto de teste para o modelo StanfordNER.

```
Evolução HNSC: NEUROLOGIA HAS = Tabagista (2 carteirasdia)= GotaATUAL:
Paciente com quadro de deficit súbito de força em hemicorpo ddireito, há 2
dias antes de internação. Relat queixa de palpitações intermitentes.=> TCC
da chegada laudo normal => TCC de controle: lacuna capsula interna esq =>
ECG : Ritmo sinusal . PR curto . Sugestivo de: Sobrecarga de cavidades
esquerdas .=> ECG 2011: fibrilação atrial => Ecocardio : AE: 41  FE:49 %=>
Labs: Sorologias NR , B12ok, Perfill Lip ok => Eco vasos cervicais : sem
estenoses  significativasPaciente  sem  queixas,  sem  registro  de
intercorrências. Ao exame: BEG , LOC , MUCPA : 16090 mmHgAC : RR, 2T,
BNFAP : MV+ s RAAB : depressivel , indolorExtremidades aquecidas e
perfundidasNEURO: sem alt pupilar , sem alt de MOE , sem alt de
sensibilidade facial , assimentria facial , deficit motor proporcionado a
dir FG 4, sem alt sesnitiva ou de coordenalçãoImpressão: AVCi de provável
TOAST lacunar, no entanto ECG 2011 com FAIRC - Cr basal : 1,8Conduta:Aguarda
holter e RM cr-nio sem gadolíneo. Ajusto dose de hidralazina e de SF .
Demais mantida. R3 Neuro [Profissional]  Dr. [Profissional] Evoluído por:
CRM em 100817 às 15:07

Solicitação de Exames HNSC: lab CRE CREATININA lab K POTASSIO lab NA SODIO
lab URE UREIA lab HEM HEMOGRAMA lab PLA PLAQUETAS CONTAGEM DE Por : CRM em
050217 às 11:42

Evolução HNSC:  Paciente  em  condiçõs  de  alta.  Amputação  ante  pé
esquerdoSegundo equipe, sem familiares.Avalio Evoluído por: CRESS em 280317
às 7:00
```

Fonte: Elaborado pelo autor.

Antes de avaliar qualitativamente o texto é preciso considerar alguns fatores inerentes a escrita em prontuários clínicos no contexto da saúde. Devido a alta demanda dos profissionais da saúde, são comuns a utilização de abreviações nos textos, o que se torna um desafio para o PLN devido a variabilidade destas abreviações. Outro ponto importante, são os erros de grafia que podem comprometer, por exemplo, o processo de tokenização e dificultar a identificação de entidades. Por fim, algumas frases carecem de estrutura semântica padrão, o que pode dificultar o processo de compreensão do texto.

Quando analisada a Figura 33 e comparadas as palavras suprimidas, é possível perceber, em relação a Figura 32, que o modelo possui uma dificuldade na identificação de algumas palavras com erros de grafias. Por exemplo, na primeira evolução as palavras **AVCi**, na linha 13, e **cr-nio**, na linha 14, deveriam ser suprimidas. Contudo, devido a erros de grafia elas não foram identificadas pelo modelo. Outro ponto que é comprometido, provavelmente, pelo erro de digitação está presente na palavra **B12ok** na linha 6. A B12 é uma vitamina, e que neste contexto está relacionada, possivelmente, com um exame de sangue.

Figura 33 – Exemplo de campos de Evolução do conjunto de teste para o modelo StanfordNER.

```
Evolução HNSC: NEUROLOGIA XXX = XXXXXXXXX (2 carteirasdia)= GotaATUAL:
Paciente com quadro de deficit súbito de força em hemicorpo ddireito, há 2
dias antes de internação. Relat queixa de palpitações intermitentes.=> XXX
da chegada laudo XXXXXX => XXX de controle: XXXXXX XXXXXXX XXXXXXX XXX =>
XXX : XXXXX XXXXXXX . XX XXXXX . Sugestivo de: XXXXXXXXXX XX XXXXXXXXX
XXXXXXXXX .=> XXX 2011: XXXXXXXXXX XXXXXX => XXXXXXXXX : XXX XX  XXXXX %=>
Labs: XXXXXXXXXX XX , B12ok, XXXXXXX XXX XX => XXX XXXXX XXXXXXXXX : sem
estenoses  significativasPaciente  sem  queixas,  sem  registro  de
intercorrências. Ao exame: XXX , XXX , XXXXX : XXXXX XXXXXX : RR, 2T,
XXXXX : XXX X XXXX : XXXXXXXXXX , indolorExtremidades aquecidas e
perfundidasNEURO: XXX XXX XXXXXXX , XXX XXX XX XXX , XXX XXX XX
XXXXXXXXXXXX XXXXXX , XXXXXXXXXX XXXXXX , XXXXXX XXXXX proporcionado a
dir FG 4, XXX XXX XXXXXXXXX ou de coordenalçãoImpressão: AVCi de provável
XXXXX lacunar, no entanto XXX 2011 com FAIRC - XX XXXXX : 1,8Conduta:Aguarda
XXXXXX e XX cr-nio sem gadolíneo. Ajusto dose de XXXXXXXXXX e de XX .
Demais mantida. R3 Neuro [Profissional]  Dr. [Profissional] Evoluído por:
CRM em 100817 às 15:07

Solicitação de Exames HNSC: lab XXX XXXXXXXXX lab K XXXXXXXX lab XX XXXXX
lab XXX XXXXX lab XXX XXXXXXXXX lab XXX XXXXXXXXX CONTAGEM DE Por : CRM em
050217 às 11:42

Evolução HNSC:  Paciente  em  condiçõs  de  alta.  Amputação  ante  pé
esquerdoSegundo equipe, sem familiares.Avalio Evoluído por: CRESS em 280317
às 7:00
```

Fonte: Elaborado pelo autor.

Contudo, de maneira geral e dada a complexidade do processamento de textos da saúde o modelo apresentou um desempenho satisfatório, conforme já validado quantitativamente. A abreviação **HAS** que neste contexto se referente a Hipertensão Arterial Sistêmica, foi identificada pelo modelo, conforme a linha 1. Esta característica, demonstra, ainda que em um contexto muito específico, que existe a possibilidade de um modelo de PLN identificar entidades com abreviações em prontuários eletrônicos.

6.3 Discussão

Para uma avaliação adequada do modelo apresentado nesse estudo, foi proposto realizar uma discussão entre os trabalhos relacionados ao tema DLPs com relevância na literatura. Dessa forma, essa seção se destina a discutir os resultados alcançados pelo modelo ARTERIAL e compara-los levando em consideração a literatura pertinente. O modelo proposto tem em termos gerais o objetivo de identificar conteúdos sensíveis e realizar uma ação específica para a proteção da informação. O estudo em questão propôs três abordagens distintas para a extração da informação, todas descritas na Seção 4.3.2.

Das abordagens testadas, duas fazem uso de redes neurais recorrentes, a seguinte utiliza algoritmo de aprendizado de máquina. Segundo Wang et al. e Vinodhini e Chandrasekaran as abordagens que levam em consideração as RNA apresentam melhores resultados na mineração de textos. Dessa forma, foram propostas abordagens utilizando LSTM e SpaCy que implementam RNR para entender e extrair informação. Dos artigos apresentados no Capítulo

3, nenhum faz uso de RNA em suas propostas. Além disso, foi proposta uma abordagem de classificação da informação utilizando algoritmo de ML, visto que foi a abordagem mais utilizada para a tarefa até o momento do levantamento dos trabalhos relacionados (HUANG et al., 2018; GONZALEZ-COMPEAN et al., 2019; ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013; ZARDARI; JUNG, 2016; THORLEUCHTER; POEL, 2012).

Levando em consideração os melhores resultados, as abordagens que fazem o uso de redes neurais apresentaram uma precisão de 87.27% e 13.0%, SpaCy e LSTM respectivamente. Já a abordagem StanfordNER que faz uso de algoritmo de ML apresentou uma precisão de 98.0%. A abordagem LSTM foi capaz de identificar com uma precisão excelente as entidades denominadas "Gerais", no entanto foi pouco eficiente em identificar as demais entidades e seus respectivos rótulos.

Dessa forma, é possível entender que o desbalanceamento na quantidade de entidades anotadas no conjunto de dados acarretou na performance insatisfatória dessa abordagem. Já os demais modelos, SpaCy e StanfordNER, possuem uma *engine* desenvolvida para o processamento de linguagem natural com configurações otimizadas, incluindo o idioma Português-br. Portanto, concluímos que essas configurações especializadas justificam o melhor desempenho do SpaCy em relação a LSTM, uma vez que ambas fazem uso de redes neurais recorrentes.

Na Tabela 12 são apresentados os resultados obtidos pelos estudos relacionados. Muitos dos artigos relacionados utilizam métricas que divergem das avaliadas no estudo em questão (GONZALEZ-COMPEAN et al., 2019; MASHECHKIN et al., 2015; KATZ; ELOVICI; SHAPIRA, 2014; WANG; JIN, 2011; THORLEUCHTER; POEL, 2012). Para uma melhor avaliação dos resultados entre os modelos de comparação, gostaríamos de olhar para a métrica F1-Score. Como a métrica F1-Score é uma média ponderada entre outros dois indicadores (Precisão e Recall), consideramos essa a melhor escolha.

Tabela 12 – Apresentação dos resultados obtidos pelos trabalhos relacionados.

Autor	Algoritmo	Acurácia	Precisão	Recall	F1-Score
(ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013)	Custom		91.4%	90.6%	
(HUANG et al., 2018)	LPA	100.0%		100.0%	
(ZARDARI; JUNG, 2016)	TsF-kNN	68.0%	47.0%	71.0%	55.0%
(DESHPANDE et al., 2015)	K-Safety		66.4%	100.0%	

Fonte: Elaborado pelo autor.

Das abordagens propostas nesse estudo, o melhor resultado foi alcançado pelo modelo StanfordNER com um F1-Score de 91.0%. Dos estudos relacionados, o melhor resultado é apresentado por Zardari e Jung com um F1-Score de 55.0%. Como apenas um dos estudos

relacionados apresenta em seus resultados o indicador F1-Score, é então analisado também o indicador Precisão. Os autores Alneyadi, Sithirasenan e Muthukkumarasamy propõem um mecanismo para a classificação de segurança das informações. Com o intuito de alcançar seu objetivo o autor define categorias, seleciona documentos, realiza o pré-processamento, extrai n-grams e implementa a classificação baseada na distância dos n-grams. Como resultado, o estudo apresenta uma Precisão de 91.4%. Dos modelos propostos nesse estudo, o StanfordNER apresenta o melhor resultado com uma Precisão de 98.0%.

Tabela 13 – Idiomas abordados pelos trabalhos relacionados.

Autor	Idioma
(ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013)	Não define
(HUANG et al., 2018)	Não define
(GONZALEZ-COMPEAN et al., 2019)	Não define
(MASHECHKIN et al., 2015)	Inglês
(KATZ; ELOVICI; SHAPIRA, 2014)	Inglês
(WANG; JIN, 2011)	Não define
(ZARDARI; JUNG, 2016)	Inglês
(THORLEUCHTER; POEL, 2012)	Alemão
(DESHPANDE et al., 2015)	Não define

Fonte: Elaborado pelo autor.

O modelo ARTERIAL faz uso do processamento de linguagem natural em seus processos. Uma definição pertinente quanto ao uso de PLN é a definição e configuração de um idioma. Conforme apresentado na Tabela 13, os estudos relacionados divergem entre o idioma Inglês, Alemão e Não definido. O idioma foi descrito como não definido quando em seu texto o autor não informou explicitamente o idioma objetivado pela proposta. Mesmo que o conjunto de dados tenha sido informado, Tabela 14, cada *dataset* disponibiliza conteúdos em mais de um idioma. Assim, não é possível afirmar o idioma treinado e avaliado. O modelo ARTERIAL tem sua proposta baseada no idioma Português-br. Dos artigos relacionados, o presente estudo é o único a abordar o referido idioma.

Outro ponto relevante a ser discutido são os conjuntos de dados utilizados. A Tabela 14

Tabela 14 – Conjuntos de dados utilizados pelos estudos relacionados

Autor	Dataset
(ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013)	Custom
(HUANG et al., 2018)	Reuters
(GONZALEZ-COMPEAN et al., 2019)	Medline
(MASHECHKIN et al., 2015)	Enron
(KATZ; ELOVICI; SHAPIRA, 2014)	Enron, Pan-PC-11, Reuters
(WANG; JIN, 2011)	Custom
(ZARDARI; JUNG, 2016)	Custom
(THORLEUCHTER; POEL, 2012)	Custom
(DESHPANDE et al., 2015)	TPC-DS

Fonte: Elaborado pelo autor.

aponta os trabalhos relacionados e os respectivos *datasets* utilizados. O *dataset* Reuters é um conjunto de documentos publicados através da sua *newswire* no ano de 1987 (LEWIS, 1997) de assuntos diversos. Já o conjunto de dados Enron fornece um grande número de mensagens de e-mail relacionados ao ambiente corporativo (KLIMT; YANG, 2004). O *PAN plagiarism corpus 2011* (Pan-PC-11) é um corpus para avaliação de algoritmos de detecção automática de plágio, gratuito para pesquisas (POTTHAST et al., 2011). O modelo ARTERIAL tem como escopo de atuação prontuários médicos eletrônicos, tornando relevante a utilização de um conjunto de dados especializado. O conjunto de dados utilizado na presente proposta é descrito na Seção 5.2.1. Levando em consideração os resultados apresentados e discutidos nessa seção, é possível concluir que o modelo proposto nesse estudo apresentou excelentes resultados em relação aos trabalhos relacionados.

7 CONCLUSÃO

O mundo está cada vez mais acessível e compartilhado, seja voltado ao pessoal ou comercial. Essa conectividade proporciona um ambiente de comunicação capaz de gerar muita informação. Ao longo dos últimos anos, a sociedade vem descobrindo o valor que a informação contém. Com o aumento na geração e compartilhamento de informação um cuidado se faz necessário. Empresas vem investindo cada vez mais em produção de informação, dedicam setores específicos a esse propósito. Como exemplo, temos os setores de pesquisa e desenvolvimento. Como complemento, novos investimentos são feitos em proteção, como tecnologias de IDS/IPS, controles de acesso, firewalls e Sistemas de DLP.

Na linha dessas tecnologias de proteção da informação, Sistemas de DLPs são responsáveis por analisar dados, realizar a compreensão e proporcionar a oportunidade de ação. Essa ação pode ser um alerta, um bloqueio de acesso ou um ocultamento da informação. No entanto, Sistemas de DLP geralmente analisam os dados a partir do contexto, não analisando com a devida profundidade o conteúdo. Além disso, como visto no Capítulo 3 que elenca os trabalhos relacionados a esse estudo, a literatura não implementa tecnologias como RNA e não fazem uso de PLN para a análise de textos não estruturados.

Dessa forma, o modelo ARTERIAL procurou analisar e aplicar tecnologias que não são utilizadas pelos modelos atuais. Assim, o modelo proposto buscou ser mais assertivo na extração da informação. A construção do modelo se deu através do acesso a uma base de dados de RES com informações não estruturadas. Esses dados foram anotados de acordo com os padrões pertinentes da área da saúde. Por conseguinte, foi realizado o pré-processamento para organização dos dados. Então três abordagens para a extração da informação foram implementadas e avaliadas, foram elas LSTM, SpaCy e StanfordNER. Como melhor resultado, o modelo StanfordNER atingiu 98.0% de Precisão, 86.0% de Recall e 91.0% de F1-Score. Enquanto isso, dos trabalhos relacionados o melhor desempenho foi apresentado pelos autores Zarrdari and Jung com 47.0%, 71.0% e 55.0%, sendo Precisão, Recall e F1-Score respectivamente. Podemos concluir assim que o modelo proposto apresentou um desempenho satisfatório em consideração aos trabalhos relacionados.

O presente estudo tinha como uma das suas principais contribuições a avaliação da utilização de RNA para a extração da informação e entidades em DLPs. Através dos experimentos realizados, foi possível avaliar o desempenho de algoritmos de aprendizado de máquina em relação a aplicação de RNA para essa tarefa. O modelo LSTM apresentou um desempenho muito divergente dos demais modelos, com um F1-Score de 14.0%. Como discutido na Seção 6.3, o modelo não foi capaz de identificar com eficiência outras entidades além da Geral. Já o modelo SpaCy, que também faz uso de RNA, apresentou um desempenho próximo aos demais modelos com um F1-Score de 80.92%. Assim, é possível concluir que RNAs apresentam melhores resultados para a extração da informação quando expostas a uma quantidade significativa de dados. Além disso, modificações especializadas podem ser

adicionadas para refinar seu funcionamento, como faz o SpaCy. Para um conjunto menor de dados, como o apresentado no conjunto de dados utilizados nesse estudo, algoritmos de aprendizado de máquina tem uma performance mais eficiente.

7.1 Trabalhos Futuros

O modelo ARTERIAL apresenta diversos pontos a serem explorados e melhorados. Como discutido na avaliação qualitativa do modelo (Seção 6.2), um ponto pertinente dos dados são os erros de grafia. Realizar a extração de entidades passa pela detecção da transformação dos dados e o erro de grafia afeta diretamente essa análise. Um mecanismo capaz de tratar esses erros pode contribuir para aprimorar a eficiência do modelo.

Outro ponto de melhoria está na tokenização dos dados na fase de pré-processamento. Como um campo não estruturado de escrita livre, muitos erros de digitação são identificados. Um erro comum encontrado pelo modelo são palavras escritas sem a utilização de espaço entre elas. Esse erro faz com que *tokens* acabem contendo mais de uma palavra, assim acarretando em prejuízo para o desempenho do modelo.

Por fim, um conjunto mais robusto de dados pode proporcionar melhor desempenho e qualidade ao modelo. Assim, qualificar e expandir o conjunto de dados utilizados nesse estudo está definido como um trabalho futuro. O conjunto pode ser qualificado através da validação por profissionais da saúde, e expandido através da anotação de mais campos livres em prontuários médicos.

7.2 Publicações

Ao longo da pesquisa, como contribuições parciais, foram produzidos diversos artigos para publicação em periódicos e eventos. A seguir são elencados os artigos que foram submetidos para avaliação:

- Trabalhos publicados durante o período de mestrado:
  - DA ROSA RIGHI, RODRIGO ; GOLDSCHMIDT, GUILHERME ; KUNST, RAFAEL ; DEON, CÁSSIO; ANDRÉ DA COSTA, CRISTIANO. **Towards combining data prediction and internet of things to manage milk production on dairy cows**. COMPUTERS AND ELECTRONICS IN AGRICULTURE, v. 169, p. 105156, 2020.
  - DA ROSA RIGHI, RODRIGO; LIMA SANTANA, ALEXANDRE; ANDRE DA COSTA, CRISTIANO; KUNST, RAFAEL; GOLDSCHMIDT, GUILHERME; KIM, DHANANJAY; SINGH, MADHUSUDAN. **Reducing Cost and Time-to-Market on Supporting Driver Assistance Systems to Avoid Rear-end Collisions in Vehicles Traffic**. In: 2019 IEEE International Conference on

Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2019, New York. 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), p. 367, 2019.

– IGNACZAK, LUCIANO; GOLDSCHMIDT, GUILHERME; ANDRÉ DA COSTA, CRISTIANO; DA ROSA RIGHI, RODRIGO. **Text mining in cybersecurity: A systematic literature review**. ACM Comput. Surv, v. 54, n. 7, 2021

– GOLDSCHMIDT, GUILHERME; NOBRE, JÉFERSON CAMPOS; DA ROSA RIGHI, RODRIGO; ANDRÉ DA COSTA, CRISTIANO. **Segurança da Informação na comunicação de dispositivos médicos: uma revisão quasi-sistemática**. Journal of Health Informatics, v. 11 , n. 2, 2019

REFERÊNCIAS

ACT, Accountability. Health insurance portability and accountability act of 1996. **Public law**, v. 104, p. 191, 1996.

AGGARWAL, Charu C; ZHAI, ChengXiang. **Mining text data**. [S.l.]: Springer Science & Business Media, 2012.

____. ____. [S.l.]: Springer Science & Business Media, 2012.

AHMAD, Farzana Kabir; YUSOFF, Nooraini. Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. In: . [S.l.: s.n.], 2013. p. 121–125.

ALLAHYARI, Mehdi et al. A brief survey of text mining: Classification, clustering and extraction techniques. **arXiv preprint arXiv:1707.02919**, 2017.

ALNEYADI, Sultan; SITHIRASENAN, Elankayer; MUTHUKKUMARASAMY, Vallipuram. Adaptable n-gram classification model for data leakage prevention. In: IEEE. **2013, 7th International Conference on Signal Processing and Communication Systems (ICSPCS)**. [S.l.], 2013. p. 1–8.

____. A survey on data leakage prevention systems. **Journal of Network and Computer Applications**, Elsevier, v. 62, p. 137–152, 2016.

ANGIANI, Giulio et al. A comparison between preprocessing techniques for sentiment analysis in twitter. In: **KDWeb**. [S.l.: s.n.], 2016.

APPARI, Ajit; JOHNSON, M Eric. Information security and privacy in healthcare: current state of research. **International journal of Internet and enterprise management**, Inderscience Publishers, v. 6, n. 4, p. 279–314, 2010.

ARCHER, Norman et al. Personal health records: a scoping review. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 4, p. 515–522, 2011.

BIRD, Steven. Nltk: the natural language toolkit. In: **Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions**. [S.l.: s.n.], 2006. p. 69–72.

BOTTOU, Léon. **Stochastic Gradient Descent Tricks**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. 421–436 p. Disponível em: <https://doi.org/10.1007/978-3-642-35289-8_25>.

BRASIL. LEI Nº 13.709 - lei geral de proteção de dados pessoais (LGPD). **Secretaria-Geral**, 2018.

BRERETON, Pearl et al. Lessons from applying the systematic literature review process within the software engineering domain. **Journal of Systems and Software**, v. 80, n. 4, p. 571 – 583, 2007. ISSN 0164-1212. Software Performance. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016412120600197X>.

BUDUMA, Nikhil; LOCASCIO, Nicholas. **Fundamentals of deep learning: Designing next-generation machine intelligence algorithms**. [S.l.]: O'Reilly Media, Inc., 2017.

CARTER, Jane V. et al. Roc-ing along: Evaluation and interpretation of receiver operating characteristic curves. **Surgery**, v. 159, n. 6, p. 1638 – 1645, 2016.

CHAPMAN, Wendy W et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. **Journal of biomedical informatics**, Elsevier, v. 34, n. 5, p. 301–310, 2001.

CHEN, Yukun et al. A study of active learning methods for named entity recognition in clinical text. **Journal of biomedical informatics**, Elsevier, v. 58, p. 11–18, 2015.

CHOLLET, Francois. **Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek**. [S.l.]: MITP-Verlags GmbH & Co. KG, 2018.

COPPIN, Ben. **Inteligência artificial**. 1. ed. Rio de Janeiro: LTC, 2010. 636 p.

DESHPANDE, Prasad M et al. The mask of zorro: preventing information leakage from documents. **Knowledge and Information Systems**, Springer, v. 45, n. 3, p. 705–730, 2015.

DU, Jie; VONG, Chi-Man; CHEN, C. L. Philip. Novel efficient rnn and lstm-like architectures: Recurrent and gated broad learning systems and their applications for text classification. **IEEE Transactions on Cybernetics**, v. 51, n. 3, p. 1586–1597, 2021.

FAWCETT, Tom. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861 – 874, 2006. ISSN 0167-8655.

FELDMAN, Ronen; SANGER, James et al. **The text mining handbook: advanced approaches in analyzing unstructured data**. [S.l.]: Cambridge university press, 2007.

FINKEL, Jenny Rose; GRENAGER, Trond; MANNING, Christopher D. Incorporating non-local information into information extraction systems by gibbs sampling. In: **Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)**. [S.l.: s.n.], 2005. p. 363–370.

FISZMAN, Marcelo et al. Automatic detection of acute bacterial pneumonia from chest x-ray reports. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 7, n. 6, p. 593–604, 2000.

FRIEDMAN, Carol et al. A general natural-language text processor for clinical radiology. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 1, n. 2, p. 161–174, 1994.

GONÇALVES, Mariana Sbaite. **Dispositivos da LGPD aplicados ao setor hospitalar**. 2021. Https://www.lgpdbrasil.com.br/dispositivos-da-lgpd-aplicados-ao-setor-hospitalar. Acesso em: 20.05.2021.

GONZALEZ-COMPEAN, JL et al. A policy-based containerized filter for secure information sharing in organizational environments. **Future Generation Computer Systems**, Elsevier, v. 95, p. 430–444, 2019.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. [S.l.]: MIT press, 2016.

____. **Deep Learning**. 1. ed. Cambridge: MIT Press, 2017.

GOYAL, Archana; GUPTA, Vishal; KUMAR, Manish. Recent named entity recognition and classification techniques: a systematic review. **Computer Science Review**, Elsevier, v. 29, p. 21–43, 2018.

GUPTA, Vishal; LEHAL, Gurpreet S et al. A survey of text mining techniques and applications. **Journal of emerging technologies in web intelligence**, v. 1, n. 1, p. 60–76, 2009.

HAYKIN, Simon. **Neural networks and learning machines**. 3. ed. New Jersey: Pearson, 2009. 938 p.

HEART, Tsipi; BEN-ASSULI, Ofir; SHABTAI, Itamar. A review of phr, emr and ehr integration: A more personalized healthcare and public health policy. **Health Policy and Technology**, Elsevier, v. 6, n. 1, p. 20–25, 2017.

HONNIBAL, Matthew et al. **spaCy: Industrial-strength Natural Language Processing in Python**. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.1212303>.

HRIPCSAK, George et al. Unlocking clinical data from narrative reports: a study of natural language processing. **Annals of internal medicine**, American College of Physicians, v. 122, n. 9, p. 681–688, 1995.

HUANG, Xiaohong et al. A novel mechanism for fast detection of transformed data leakage. **IEEE Access**, IEEE, v. 6, p. 35926–35936, 2018.

INFOWATCH. **Global Data Leakage Report 2018**. 2019. Disponível em: <https://infowatch.com/report2018_half>.

ISO/IEC:27001. **Information technology — Security techniques — Information security management systems — Requirements**. [S.l.], 2013.

ISO/TR:14639-2. **Health informatics - Capacity-based eHealth architecture roadmap - Part 2: Architectural components and maturity model**. [S.l.], 2014.

JIA, Haosen et al. The latent semantic power of labels: Improving image classification via natural language semantic. In: SPRINGER. **International Conference on Human Centered Computing**. [S.l.], 2019. p. 175–189.

JIANG, Ridong; BANCHS, Rafael E; LI, Haizhou. Evaluating and combining name entity recognition systems. In: **Proceedings of the Sixth Named Entity Workshop**. [S.l.: s.n.], 2016. p. 21–27.

JO, Taeho. **Text Mining: Concepts, Implementation, and Big Data Challenge**. [S.l.]: Springer, 2018.

JOHNSON, Alistair EW et al. Mimic-iii, a freely accessible critical care database. **Scientific data**, Nature Publishing Group, v. 3, n. 1, p. 1–9, 2016.

JURAFSKY, Dan. **Speech & language processing**. [S.l.]: Pearson Education India, 2000.

KANNAN, Subbu; GURUSAMY, Vairaprakash. Preprocessing techniques for text mining. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2014.

KATZ, Gilad; ELOVICI, Yuval; SHAPIRA, Bracha. Coban: A context based model for data leakage prevention. **Information sciences**, Elsevier, v. 262, p. 137–158, 2014.

KINGMA, Diederik P; BA, Jimmy. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.

KITCHENHAM, Barbara Ann; BUDGEN, David; BRERETON, Pearl. **Evidence-based software engineering and systematic reviews**. [S.l.]: CRC press, 2015.

KLIMT, Bryan; YANG, Yiming. The enron corpus: A new dataset for email classification research. In: SPRINGER. **European Conference on Machine Learning**. [S.l.], 2004. p. 217–226.

LEWIS, David. Reuters-21578 text categorization test collection, distribution 1.0. **http://www. research/. att. com**, AT&T Labs-Research, 1997.

LIDDY, Elizabeth D. Natural language processing. 2001.

LOZANO-RUBÍ, Raimundo et al. Ontocr: A cen/iso-13606 clinical repository based on ontologies. **Journal of biomedical informatics**, Elsevier, v. 60, p. 224–233, 2016.

MANNING, Christopher D; MANNING, Christopher D; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. [S.l.]: MIT press, 1999.

MANNING, Christopher D et al. The stanford corenlp natural language processing toolkit. In: **Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations**. [S.l.: s.n.], 2014. p. 55–60.

MASHECHKIN, IV et al. Applying text mining methods for data loss prevention. **Programming and Computer Software**, Springer, v. 41, n. 1, p. 23–30, 2015.

MCCANN, E. Kaiser reports second fall data breach. **Healthcare IT News**, 2013.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943.

MELTON, Genevieve B; HRIPCSAK, George. Automated detection of adverse events using natural language processing of discharge summaries. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 12, n. 4, p. 448–457, 2005.

MERCURI, Rebecca T. The hipaa-potamus in health care data security. **Communications of the ACM**, ACM New York, NY, USA, v. 47, n. 7, p. 25–28, 2004.

MILSTEIN, Ricarda; BLANKART, Carl Rudolf. The health care strengthening act: the next level of integrated care in germany. **Health Policy**, Elsevier, v. 120, n. 5, p. 445–451, 2016.

MOUSTAKA, Vaia; VAKALI, Athena; ANTHOPOULOS, Leonidas G. A systematic review for smart city data analytics. **ACM Computing Surveys (CSUR)**, ACM, v. 51, n. 5, p. 103, 2018.

MUNKOVÁ, Daša; MUNK, Michal; VOZÁR, Martin. Data pre-processing evaluation for text mining: transaction/sequence model. **Procedia Computer Science**, Elsevier, v. 18, p. 1198–1207, 2013.

NEWAZ, AKM et al. A survey on security and privacy issues in modern healthcare systems: Attacks and defenses. **arXiv preprint arXiv:2005.07359**, 2020.

OLIVEIRA, Douglas Nunes de; MERSCHMANN, Luiz Henrique de Campos. Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in brazilian portuguese language. **Multimedia Tools and Applications**, Springer, v. 80, n. 10, p. 15391–15412, 2021.

OUELLET, Eric; MCMILLAN, Rob. Magic quadrant for content-aware data loss prevention. **Gartner Inc**, 2013.

PAI, Madhukar et al. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. **THE NATIONAL MEDICAL JOURNAL OF INDIA**, v. 17, n. 2, p. 86–95, 2004. Disponível em: <http://archive.nmji.in/Issue%2017-2%20PDF/Systematic-reviews-and-meta-analyses.pdf>.

PATIL, Harsh Kupwade; SESHADRI, Ravi. Big data security and privacy issues in healthcare. In: IEEE. **2014 IEEE international congress on big data**. [S.l.], 2014. p. 762–765.

PEDRO, Ricardo Wandré Dias; MACHADO-LIMA, Ariane; NUNES, Fátima L.S. Is mass classification in mammograms a solved problem? - a critical review over the last 20 years. **Expert Systems with Applications**, v. 119, p. 90 – 103, 2019.

PONEMON, Institute Research Report. Benchmark study on patient privacy and data security. **Journal of healthcare protection management: publication of the International Association for Hospital Security**, v. 27, n. 1, p. 69, 2011.

POTTHAST, Martin et al. **PAN Plagiarism Corpus 2011 (PAN-PC-11)**. Zenodo, 2011. Disponível em: <https://doi.org/10.5281/zenodo.3250095>.

RAJMAN, Martin; VESELY, Martin. From text to knowledge: Document processing and visualization: A text mining approach. In: **Text mining and its applications**. [S.l.]: Springer, 2004. p. 7–24.

RAMAN, Preeti; KAYACIK, Hilmi Güneş; SOMAYAJI, Anil. Understanding data leak prevention. In: CITESEER. **6th Annual Symposium on Information Assurance (ASIA'11)**. [S.l.], 2011. p. 27.

RINGGER, Eric et al. Active learning for part-of-speech tagging: Accelerating corpus annotation. In: **Proceedings of the Linguistic Annotation Workshop**. [S.l.: s.n.], 2007. p. 101–108.

ROBBINS, Stanley L; COTRAN, Ramzi S; KLATT, Edward C. **Robbins and Cotran Atlas of Pathology**. [S.l.]: Elsevier, 2015.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial**. Rio de Janeiro: Elsevier, 2013. 988 p.

RUUSKA, Salla et al. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. **Behavioural Processes**, v. 148, p. 56 – 62, 2018. ISSN 0376-6357.

SCHMITT, Xavier et al. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In: **2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)**. [S.l.: s.n.], 2019. p. 338–343.

SECUROSIS, LLC. Understanding and selecting a data loss prevention solution. In: **Securosis**. [S.l.]: LLC, 2010.

SERVICES, Centers for Medicare & Medicaid et al. **Health Information Technology for Economic and Clinical Health Act-Electronic Health Record Incentive Program; Final Rule. 2010**. 2011. Disponível em: <http://edocket.access.gpo.gov/2010/pdf/2010-17207.pdf>.

SHABTAI, Asaf et al. Detecting unknown malicious code by applying classification techniques on opcode patterns. **Security Informatics**, SpringerOpen, v. 1, n. 1, p. 1, 2012.

SHU, Xiaokui; YAO, Danfeng Daphne. Data leak detection as a service. In: SPRINGER. **International Conference on Security and Privacy in Communication Systems**. [S.l.], 2012. p. 222–240.

SINGH, Jasmeet; GUPTA, Vishal. A systematic review of text stemming techniques. **Artificial Intelligence Review**, Springer, v. 48, n. 2, p. 157–217, 2017.

SLANKAS, J. et al. Relation extraction for inferring access control rules from natural language artifacts. In: **Proceedings of the 30th Annual Computer Security Applications Conference**. Association for Computing Machinery, 2014. (ACSAC '14, December), p. 366–375. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954538008&doi=10.1145%2f2664243.2664280&partnerID=40&md5=62364fff2be378a947d497a222989945>.

SOYSAL, Ergin et al. Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 25, n. 3, p. 331–336, 2018.

STANDARDIZATION, International Organization for. **Information technology-Security techniques-Code of Practice for Information Security Management**. [S.l.]: na, 2005.

SULLIVAN, Frost &. **World Data Leakage Prevention Market**. United States, 2008.

TAHBOUB, Radwan; SALEH, Yousef. Data leakage/loss prevention systems (dlp). In: IEEE. **2014 World Congress on Computer Applications and Information Systems (WCCAIS)**. [S.l.], 2014. p. 1–6.

TEAM, Verizon RISK et al. Data breach investigations report. **Retrieved May**, v. 7, p. 2013, 2013.

THORLEUCHTER, D.; POEL, D. Van den. Improved multilevel security with latent semantic indexing. **Expert Systems with Applications**, Pergamon-elsevier Science Ltd, v. 39, n. 18, p. 13462–13471, dez. 2012. ISSN 09574174. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84865269030&doi=10.1016%2fj.eswa.2012.06.002&partnerID=40&md5=4f8ec1bee2d07bc913e77a97a991dd79>.

TIAN, Zhihong et al. Deep learning and dempster-shafer theory based insider threat detection. 2019.

TRAN, Huy; ZDUN, Uwe et al. Systematic review of software behavioral model consistency checking. **ACM Computing Surveys (CSUR)**, ACM, v. 50, n. 2, p. 17, 2017.

UYSAL, Alper Kursat; GUNAL, Serkan. The impact of preprocessing on text classification. **Information Processing & Management**, Elsevier, v. 50, n. 1, p. 104–112, 2014.

UZUNER, Özlem et al. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 552–556, 2011.

VERMA, Yugesh. **Complete Guide To Bidirectional LSTM**. 2021. Disponível em: <https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/>.

VINODHINI, G; CHANDRASEKARAN, RM. A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. **Journal of King Saud University-Computer and Information Sciences**, Elsevier, v. 28, n. 1, p. 2–12, 2016.

VYCHEGZHANIN, Sergey; KOTELNIKOV, Evgeny. Comparison of named entity recognition tools applied to news articles. In: **2019 Ivannikov Ispras Open Conference (ISPRAS)**. [S.l.: s.n.], 2019. p. 72–77.

WANG, Peng et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. **Neurocomputing**, Elsevier, v. 174, p. 806–814, 2016.

WANG, Qihua; JIN, Hongxia. Data leakage mitigation for discretionary access control in collaboration clouds. In: ACM. **Proceedings of the 16th ACM symposium on Access control models and technologies**. [S.l.], 2011. p. 103–112.

WEIGUO, Fan et al. Tapping into the power of text mining. **Journal of ACM, Blacksburg**, 2005.

WEISS, Sholom M et al. **Text mining: predictive methods for analyzing unstructured information**. [S.l.]: Springer Science & Business Media, 2010.

YANG, Jianwu; CHEN, Xiaoou. A semi-structured document model for text mining. **Journal of Computer Science and Technology**, Springer, v. 17, n. 5, p. 603–610, 2002.

ZARDARI, Munwar Ali; JUNG, Low Tang. Data security rules/regulations based classification of file data using tsf-knn algorithm. **Cluster Computing**, Springer, v. 19, n. 1, p. 349–368, 2016.

ZEISER, Felipe André et al. Deepbatch: A hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images. **Expert Systems with Applications**, Elsevier, v. 185, p. 115586, 2021.

ZEISER, Felipe André et al. Breast cancer intelligent analysis of histopathological data: A systematic review. **Applied Soft Computing**, v. 113, p. 107886, 2021. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494621008085>.

ZHAI, ChengXiang; MASSUNG, Sean. **Text data management and analysis: a practical introduction to information retrieval and text mining**. [S.l.]: Morgan & Claypool, 2016.

ZHANG, Shaodian; ELHADAD, Noémie. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. **Journal of biomedical informatics**, Elsevier, v. 46, n. 6, p. 1088–1098, 2013.